Strategic Competence, Task Complexity, and Learner Performance in Computer-delivered Integrated Speaking Test Tasks: A Study of English-as-a-Foreign-Language (EFL) Learners in China

Weiwei Zhang

A thesis submitted in fulfilment of the requirements for
the Degree of Doctor of Philosophy (Ph.D.) in Education
(Applied Linguistics & TESOL)

The University of Auckland, Auckland
New Zealand

2021

Abstract

Understanding the relationships among test-takers' strategic competence, test tasks and test performance is a perennial problem in language assessment. Despite numerous research efforts in addressing the problem, little is known about how the intricate relationships play out in the integrated speaking tests, one of the most popular formats in high-stakes tests such as the computer-mediated TOEFL (the Test of English as a Foreign Language). This PhD study aims to fill the research gap in a Chinese EFL context.

Underpinned by Bachman and Palmer's (2010) framework of non-reciprocal language use, the study employed a convergent mixed-methods design involving 616 Chinese EFL students and five Chinese EFL teachers to examine whether Chinese EFL learners' speaking performance was affected by their strategic competence (i.e., their use of metacognitive strategies) and task complexity. It also investigated if the interaction between strategic competence and task complexity would affect speaking performance. To this end, the student participants answered a questionnaire to report their metacognitive strategy use immediately after they performed four TOEFL-based integrated speaking test tasks. All the participants self-rated the difficulty of the test tasks to measure task complexity. Eight students participated in the subsequent interviews for an in-depth probe into the Chinese EFL learners' metacognitive strategy use and their perceptions of task complexity. To investigate the intricate relationships, statistical procedures such as one-way repeated-measures MANOVA, structural equation modelling, one-way repeated measures ANOVA, hierarchical linear modelling, and multiple regression were used at different stages of quantitative data analysis. Qualitative coding with content analysis was also conducted.

The findings suggest that among the four individual subcomponents of the metacognitive strategies (planning, problem-solving, monitoring and evaluating), only problem-solving reported by the Chinese EFL learners demonstrated substantial variance across the speaking test tasks. By contrast, these learners' use of interactive metacognitive strategies and their speaking performance were significantly affected by task complexity. Regarding the relationships among the three variables, monitoring as

a metacognitive strategy moderated the effect of task complexity on the learners' speaking performance. In the same vein, prior knowledge was identified as a moderator between the other three task complexity variables (viz., planning time, steps involved and task type) and the learners' overall perceptions of task complexity involved in the test tasks. Unexpectedly, individual attributes (e.g., motivation and anxiety), though not the focus of the study, were found to mediate the interaction between the learners' metacognitive strategy use and task complexity.

The findings primarily lend some new evidence in support of Bachman and Palmer's (2010) Strategic Competence Model and their proposed framework of non-reciprocal language use. They also support Robinson's (2015) Triadic Componential Framework and Kormos' (2011) Bilingual Speech Production Model. Taken together, these findings provide implications for metacognitive scaffolding, syllabus designing and task development for EFL speaking instructions. Moreover, the findings are expected to add empirical evidence for validity arguments for test development: In designing speaking tests, there appears to be a need for taking into consideration task complexity so that the tests can truly assess test-takers' language ability, as required for meeting the assumptions of test validity and reliability.

Acknowledgements

I always believe my doctoral study is a tough journey, and without the supervision and support from my supervisors, Professor Lawrence Zhang and Associate Professor Aaron Wilson, I would have never arrived at my destination. Thanks to their company all the way with their expertise, patience, encouragement and inspirations, I finally completed the first purely academic adventure in my life with a relief. They are the ones to whom I should express my deepest appreciation and gratitude.

I would also like to say a word of thanks to Professor Yu Jianhua and Professor Zhao Meijuan from Shanghai International Studies University, without whom I was not able to have the chance of starting the journey.

The email correspondences and the face-to-face talks on statistics with Professor Li Chao from Anhui University of Finance and Economics, Professor Gavin Brown and Dr. Kane Meissel from The University of Auckland, and Professor Wen Fuxing from Soochow University in Taipei have turned into the numbers, figures, and words that provide powerful evidence for this thesis. The frequent chatting with my fellow doctoral student Joshua Sarpong gave me not only fresh ideas on my research, but invaluable friendship that I will keep in mind for long. The participation of the students and the teachers and the support of the leaders from the universities enabled me to lift up a research proposal to a PhD thesis.

I need to record my appreciation to the English Testing Service (ETS) USA for awarding me with a small doctoral research grant, which rendered some funding for this research. My main supervisor's support to me for applying for this small grant is tremendously important, which I also need to duly acknowledge.

Finally, I want to say "thank you" with a warm hug to my family for their all-around support, love, and care throughout.

Dedication

In memory of the special times in the COVID-19 lockdown.

"It was the best of times, it was the worst of times…"

— Charles Dickens

When the COVID-19 was torturing the whole world, forcing us to live in lockdown and isolation, I finished the first draft of this thesis earlier than expected. It is the isolation that drove me to stay far away from the hustle-and-bustle, to read more, and to think more. When we are mourning for those who lost their lives or sympathising with those who lost their beloved ones in the disaster of our humankind, we have learned who we are and where we are going.

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction

1.1. Chapter Overview

This chapter sets out to provide a holistic view of the study. It begins with a presentation of the research aims and proceeds to discuss the motivation through a brief review of the relevant literature, which points to the research gaps that the present study is expected to fill. This is followed by an examination of the research context in which the actuality of learning and teaching EFL speaking in China is introduced. The discussion about the motivation and the research context justify the significance of this study. It then moves on to the scope, the objectives, and the research questions before closing with an introduction to the organisation of this thesis.

1.2 Aims

Over a decade ago, Bachman (2007) cautioned us of the challenge facing language assessment. He posited: "A persistent problem in language assessment has been that of understanding the roles of abilities and contexts, and the interactions between these, as they affect performance on language assessment tasks" (p. 1). Unfortunately, this still remains as a challenge (Hughes & Reed, 2017), especially in L2 speaking assessment, a field that is under-researched (Fulcher, 2015a, 2015b). Against this backdrop, this study investigated the relationships among strategic competence of the English-as-a-Foreign Language (EFL) learners in China, their speaking performance, and task complexity involved in computer-delivered integrated speaking test tasks. It aims to enrich our understandings of language assessment that help to address the problem by providing additional empirical evidence from a Chinese perspective. It is also hoped that the findings can contribute to the current literature on language assessment by adding empirical evidence to the validation of the framework of non-reciprocal language use and of the strategic competence model in relation to language ability (Bachman & Palmer, 2010). Moreover, the probe into the test tasks within Robinson's (2015) Triadic Componential Framework and the speaking process in Kormos' (2011) Bilingual Speech Production Model is expected to provide additional validation support for the

framework and the model, and accordingly promote the development of learning and teaching EFL speaking in terms of task sequencing for pedagogic practice and test development.

1.3 Motivation

As Creswell and Guetterman (2019) stated, research can be built upon one's workplace or personal experiences, the initial impetus of this study was my working experience in helping Chinese EFL learners to prepare for the Testing of English as Foreign Language (TOEFL) speaking, a computer-mediated integrated speaking test.

1.3.1 Statement of the problem

As an English teacher, previously, my job was to teach speaking to a special group of Chinese EFL learners: Candidates matriculated into cultural exchange programmes for their further studying abroad. Because entry into such a programme was highly competitive in English proficiency, on average, my students were high-proficiency EFL learners. Also, they had to pass international English language tests such as TOEFL prior to their admission to tertiary studies overseas. Despite this, in classroom assessment, most performed worse on the integrated speaking tasks where reading, listening, and speaking skills are integrated than on independent speaking tasks where only speaking is involved. More confusingly, those "weak" students (students with relatively lower levels of language proficiency) occasionally outperformed their peers on the integrated speaking tests. The confusing phenomenon prompted me to ask the following questions: What might account for my students' poor performance on the integrated speaking tests? Is it their language proficiency, the integrated speaking test or both? If so, how do these factors affect the students' performance?

In my extensive review of the published language assessment research, I found that test-takers' language proficiency is also known as their language ability, which works independently or interactively with test tasks, affecting test performance (e.g., Hughes, 2017; Hughes & Read, 2017; O'Sullivan, 2019; see also earlier significant work of Bachman & Palmer, 2010; Luoma, 2004; Weir, 2005). In line with this, my students' test performance essentially reflected the interaction between their language ability and the test task characteristics involved in the integrated speaking tests. As such, to improve

my students' test performance, it is necessary to understand test-takers' language ability, test task characteristics and the interaction between them.

In modern language assessment, researchers tend to define test-takers' language ability within a specific model or framework (e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale & Swain, 1980; Lado, 1961; Oller, 1983). The independent application of these models and frameworks in examining language assessment, however, has failed to duplicate the interactional properties involved in a language test, as pointed out by scholars such as Bachman (2007), Bachman and Palmer (2010), and Luoma (2004), among others. To address this issue, Bachman and Palmer (2010) proposed the conceptualisation of language ability in an interactional framework of language use in parallel with a framework of test task characteristics. Based on the proposal, they further put forward two frameworks: The framework of reciprocal language use for inter-individual interactions and the framework of non-reciprocal language use for situations where individual-to-individual conversations are not required, such as computer-assisted language speaking tests (e.g., TOEFL). It is obvious that the non-reciprocal characteristic of the second framework is relevant to my context because the computer-administered integrated speaking tests have been a great challenge to my students. This encouraged me to frame my investigation into the questions from my classroom instructions within this second non-reciprocal framework. Alternatively stated, the framework of non-reciprocal language use serves as the theoretical framework of this study (see Chapter Two for the detailed discussion of the rationale).

In the framework of non-reciprocal language use, test task characteristics are illustrated in the Language Use Task Characteristics Model composed of the setting, rubrics, input, response, and the relationship between the input and the response. With reference to this model, the integrated speaking test tasks are interpreted as the characteristic of input format, which indicates the varying degrees of task complexity (Brown, Iwashita, & McNamara 2005; Cox, 2013; Getman, 2020; Liu & Li, 2012; Robinson, 2011a, 2011b, 2015; Robinson & Gilabert, 2020). On the other hand, language ability is defined by Bachman and Palmer (2010) within the Language Ability Model as consisting of language knowledge and strategic competence. Chapelle et al. (2011) interpreted the two components as individuals' knowledge and use of a language respectively. The

former concerns one's knowledge of grammar and vocabulary and the latter involves one's ability to use different strategies. In a language test, test-takers often use a combination of knowledge and strategies to complete test tasks, hence the knowledge of language per se is not enough for them to perform the tests. Chapelle and colleagues' interpretation of language ability resonates with that of Sun (2014) who used linguistic competence and strategic competence to name the two components. Sun stated that Chinese EFL learners' language ability and the EFL teaching in China are characterised by an "overdue bias on linguistic competence rather than strategic competence", which obviously portrays my students' actuality in learning EFL speaking. She also commented that "linguistic competence has always earned the greatest attention in language learning, but strategic competence still lacks due attention. Thus, striking a proper balance between these two elements has become a great challenge for foreign language teachers in China" (p. 1604). Given the roles of language knowledge and strategic competence in language assessment and the characteristics of Chinese EFL learners' language ability, it is reasonable that the focus of this study is placed on the strategic competence used by Chinese EFL learners in their actual test performance over their language knowledge to examine their language ability.

1.3.2 Research gaps

Having reviewed the literature on language assessment, I believed that the answers to the questions from my classroom instructions could be addressed through an examination of the relationships among Chinese EFL learners' strategic competence, or their metacognitive strategy use, task complexity, and their performance on the integrated speaking test tasks. My subsequent review of the literature regarding the three variables within the theoretical framework suggested that studies on metacognitive strategies, task complexity and speaking performance had tapped not only language assessment, but also the wider research fields of psychology, applied linguistics, and education. Nevertheless, empirical studies exploring their relationships are scant, which indicates the research gaps.

In Bachman and Palmer's (2010) framework of non-reciprocal language use, strategic competence is the use of metacognitive strategies. Though emanating from psychology as one of the fundamental subordinates of metacognition (Zhang & Qin, 2018; Zhang &

Zhang, 2018), metacognitive strategies are commonly perceived as a salient factor in language learning and language assessment. They work independently and interactively, affecting language performance (e.g., Chamot & Harris, 2019; Griffiths, 2013; Takeuchi, 2020, Yi, 2012). Such importance has stimulated a large volume of studies that investigate the correlations between metacognitive strategies and performance. However, these studies mainly focus on listening, reading, and writing in non-testing situations (e.g., Qin, 2018; Teng & Zhang, 2016; Zhang, 2010; Zhang, Zhang & Wu, 2009), and researchers typically examine the working of metacognitive strategies in an independent form, paying little attention to the interactive working mode of these strategies in an actual assessment context. Hence, research into how metacognitive strategies work independently and interactively in a simultaneous manner in speaking to affect language performance in authentic tests is lacking, despite the theoretical grounding that underscores the indispensable role of metacognitive strategies in speaking assessment (e.g., Barkaoui et al., 2013; Bygate, 2011; Fernandez, 2018; H. Huang, 2016; L. Huang, 2013; Hughes 2017, Luoma, 2004).

In language assessment, speaking has been accepted as a process in which the speakers' metacognitive monitoring operates both covertly and overtly for the completion of the speaking tasks. The use of strategic competence empowers speakers to plan the knowledge at hand and to compensate for and facilitate their oral production (Bygate, 2011; Cox, 2020; Kormos, 2006, 2011; Luoma, 2004; Skehan, 2016, 2018; Yi, 2012). This explains why the integrated speaking tests "broaden the scope of strategies called upon (Barkaoui et al., 2013, p. 16), and are closely related to pre-assessment and pre-planning, online planning and monitoring, and post-evaluation (Cohen, 2014). The interwoven relationship between metacognitive strategies and the speaking tasks undoubtedly forms a sharp contrast with the few research efforts exploring this topic. Such actuality justifies the call from scholars for further empirical research in L2 speaking assessment (e.g., Huang, 2016; Seong, 2014; Wigglesworth & Frost, 2017).

With regard to test task characteristics, as a form of input, the integrated speaking test tasks involve the simultaneous manipulation of various task complexity variables, simulating the real-life language use situations and contributing to the authenticity and predictive validity of the speaking tasks (e.g., Barkaoui et al., 2013; Huang, 2016; Swain

et al., 2009; Wigglesworth & Frost, 2017; Yang & Plakans, 2012). However, in the existing literature on task complexity, studies on the simultaneous variation in various task factors in speaking are relatively insufficient (see, e.g., comments by Adams & Alwi, 2014), except for those conducted by Gilabert (2005, 2007a, 2007b), Gilabert and Llanes (2009), Levkina (2008), and Levkina and Gilabert (2012). Furthermore, in addressing the impact of task complexity on speaking performance, researchers commonly use self-designed tasks rather than authentic speaking test tasks to manipulate task complexity. Consequently, whether these tasks reflect the designed task complexity is not clear (Sasayama, 2015, 2016). Due to this, the validity and the reliability of these tasks and the generalisability of the research results may not be without caveats (Bachman & Palmer, 2010; Creswell & Creswell 2018; Sasayama, 2015; Skehan, 2011, 2016, 2018).

Moreover, it is apparent that the empirical attempts discussed above are mainly concerned with the isolated investigation of the interrelationships between test-takers' metacognitive strategy use and their speaking performance, and between task complexity and test performance. This body of research has not shown the interactional features of the speaking assessment (Seong, 2014). In fact, only four studies, to my best knowledge, have examined the interactions (e.g., Barkaoui et al., 2013; Huang, 2016; Swain et al., 2009; Yi, 2012). However, in these studies, participants were just a small group of EFL learners who had exposure to an English-speaking environment. Such sample characteristic makes the research findings infeasible to be generalised to a larger population of EFL learners, especially for EFL learners in China who have no easy access to an English-speaking environment (Gu, 2014; Zhang, 2010; Zhang & Qin, 2018). Given that China is claimed to have the largest population of EFL learners in the world (Zhang & Qin, 2018), for whom speaking is the biggest hurdle (see Section 1.4), there is an urgent need for probing into how Chinese EFL learners approach speaking test tasks with varying degrees of complexity in a larger sample. Additionally, as metacognitive strategies work independently and interactively in language performance, it is warranted that a comprehensive perspective should be taken in examining the role of these strategies in individuals' actual language use for better performance (Takeuchi, 2019, Yi, 2012). The present study is an effort in this research direction where the independent and interactive metacognitive strategies, used by a larger sample of Chinese

EFL learners in the interactional speaking process as a response to the integrated speaking test tasks, are examined.

1.4 Research Context

Since this study is from, on and mainly for EFL learners in China, it is of the essence to delve into how English speaking is taught and learnt in this country.

1.4.1 Learning and teaching EFL speaking in China's educational system

In China's educational system, English is a compulsory subject that all students must learn from Grade 3 in primary school to prepare for the university entrance examination, and the fundamental objective of EFL education is to enable students to use English, especially their listening and speaking skills for effective communications (Ministry of Education, 2017). For some universities, one of the requirements for students to obtain their university certificate is to pass the English examinations to a certain level (Haidar & Fang, 2019). Because of the great importance of English, EFL learning has elicited enormous efforts in China, and the figure for Chinese EFL learners reached approximately 400 million as early as in 2015 (Wei & Su, 2015), outnumbering the combined population of the UK and the USA (Fang, 2018).

Despite the importance of speaking in China's EFL education and the great enthusiasm about English learning in the country, it is common that many Chinese EFL learners find that after many years of learning, it is hard to correctly express their ideas in an actual situation where speaking serves as the only medium of communications, (Dou, 2013; Gorsuch, 2011; Zhang, 2006). Some scholars (e.g., Pan & Li, 2012; Sun, 2014; Tian, 2012; Tubbs, 2016; Zhao, 2017) comment that the fundamental cause of this phenomenon is that Chinese EFL teachers typically emphasise students' knowledge of a language (e.g., the vocabulary and the grammar) over their strategic competence, a competence highly demanded in speaking (Bygate, 1987, 2011; Luoma, 2004; Kormos, 2011). Similarly, Chinese EFL teachers normally teach speaking, listening, writing, and reading separately rather than in an integrated manner (Cai, 2012; Pan & Li, 2012). In the real-world context, however, language skills are used interdependently in that it is impossible, logically and practically, to break the use of a language into isolated language skills, as language users must receive input in either a written or spoken form

before a real communication begins (Chen, 2020; Crossley & Kim, 2019; Brown & Ducasse; 2019). As a result, the teaching practice of the Chinese EFL teachers cannot reflect the real language use context. In fact, some scholars have advocated that EFL teachers should teach learners to integrate language skills so that learners can interact naturally in an authentic language use situation. They have also suggested that integrating language skills should be an important pedagogical component in EFL classroom instructions (e.g., Newton & Nation, 2020; Oxford, 2001). Nevertheless, teaching EFL in an integrated fashion has not been a popular pedagogic practice in China (Cai, 2012; Pan & Li, 2012).

Chinese EFL teachers' long-standing lack of attention to help their students develop strategic competence, and their pedagogic activities that are impossible to simulate the real language use situations have characterised EFL classroom instructions in China. These characteristics have placed Chinese EFL learners in a disadvantageous position on occasions when speaking is highly demanded, such as studying abroad (Cai, 2012; Li & Pan 2012; Sun, 2014; Tubbs, 2016).

1.4.2 Learning and teaching EFL speaking for Chinese learners with special need of overseas study

With the booming cooperation between China and other countries amid globalisation, the number of international schools in China reached about 1,000 in 2019, ranking the first worldwide, and the number of universities that co-hosted exchange programmes with their foreign partners arrived at 2, 626 at the beginning of 2018 (Zhang & Huang, 2019). Through these international co-operations, many Chinese EFL learners went overseas for further education. Taking the USA as an example, the latest data from the Open Door 2019 Report reveal that in the academic year of 2018/2019, the number of Chinese students enrolled in USA universities ranked the first, arriving at 368,548, which accounted for 31.5% of the country's total international students and represented a yearly increase of 8.1%. The statistics provides evidence that a huge population of Chinese EFL learners have the need to study overseas, including my students.

For Chinese EFL learners, one of the fundamental requirements for being admitted by overseas academic institutes is evidence of their English proficiency level: They must

sit international language tests such as TOEFL and get a mark requested. To help these students prepare for these examinations, educational institutes at various levels provide relevant language courses. Classroom instructions in these courses have been revealed as heavily influenced by the content and the format of these tests (Cheng, Rogers & Hu, 2004; Yu et al., 2017). Yet, due to lack of professional knowledge and unitary standards, Chinese EFL instructors of these courses do not have a standard syllabus for integrated speaking test tasks characterised by these language tests and they do not have any specific theoretical or systematic training background on the tests either (Yu et al., 2017). These teachers' pedagogical practice is usually based on their understanding of or their own experience in taking the tests. Consequently, these teachers might lack confidence, or even suffer from anxiety in classroom instructions (Tian, 2012; Zhou, 2020). This may discourage their students in EFL learning, since teachers' authority is normally emphasised as a leading factor in the Chinese classrooms (Qin, 2007; Wen, 2016). These real situations of learning and teaching EFL speaking for overseas study may explain why Chinese students "often lack the fluency and the intelligibility needed to converse comfortably in English" in American universities (Tubbs, 2016, p. 3).

Indeed, beyond the USA and within the global setting, Chinese EFL learners' speaking ability is also a major concern for many researchers. They point out that Chinese EFL learners seem to be unprepared and unqualified for their overseas study in terms of their speaking ability (e.g., Gorsuch, 2011; Tian, 2012; Tubbs, 2016; Zhou, 2020). This concern indicates why the biggest hurdle in Chinese EFL learners' preparation for TOEFL is the speaking test. According to the annual report on Chinese test-takers' TOEFL scores issued by ETS (English Testing Service) in 2019, the average score of Chinese test-takers' speaking was 19 points (the total score for each section of TOEFL is 30 points), lower than reading (21 points), listening (20 points) and writing (20 points). This figure remained unchanged for two consecutive years from 2017 to 2019.

In summary, the actuality of learning and teaching of EFL speaking in China justifies that additional empirical studies on speaking, in particular, on the integrated speaking tests, are needed to foster the capability of Chinese EFL speakers in the authentic language use domains.

1.5 Significance

The current study is expected to have theoretical, methodological, and pedagogical implications in relation to the aforementioned research niche in the existing literature, while addressing the questions from my classroom instructions.

Theoretically, this study took into account the performance of metacognitive strategies that are composed of four elements (planning, problem solving, monitoring, and evaluating) in both independent and interactive forms in the integrated speaking test tasks. As discussed earlier, very few studies have been conducted in this line. Therefore, the study is likely to add new information to test-taker's strategic competence in speaking assessment. This information will contribute to the validation of Bachman and Palmer's (2010) strategic competence model and their framework of non-reciprocal language use (e.g., Phakiti, 2016; Purpura, 2013; Zhang, 2017; Zhang, Aryadoust, & Zhang, 2014). Methodologically, the Metacognitive Strategy Inventory that was employed to examine strategic competence is original. Since there is no such an instrument targeting at the performance of strategic competence used by EFL learners, particularly, Chinese EFL learners in the context of integrated speaking tests, the development and validation of the questionnaire will partially bridge the research gap. Pedagogically, as this study involves the examination of Chinese EFL learners' metacognitive strategy use, the research findings will empower Chinese EFL teachers to facilitate their metacognitive scaffolding in classroom instructions. In addition, the investigation of task complexity of the integrated speaking test tasks may help Chinese EFL teachers understand the factors that can be manipulated to change task complexity in designing and sequencing tasks used in their syllabus. It may also inspire the teachers to teach EFL in an integrated way. In the same vein, the understanding of task complexity can help test developers develop test tasks with appropriate complexity so that the tests can truly measure test-takers' language ability for test validity and reliability.

1.6 Scope, Objectives and Research Questions

Base on the above delineation, the scope of this study is clarified. In line with the scope, the objectives and accordingly the research questions are developed as presented in the following subsections.

1.6.1 Scope

In a computer-mediated speaking test, a test-taker's performance is affected by many factors that interact with one another including test-takers' attributes, task types, or other random factors (e.g., Bachman, 1990; 2002; 2007; Bachman & Palmer, 1996; 2010; Barkaoui 2013; O'Sullivan, 2000; 2012; Purpura, 1997, 1999, 2016; Skehan, 1996; Weir, 2005). All of the factors are included in Bachman and Palmer's (2010) framework of non-reciprocal language use. Although the framework underpinned the present study theoretically, only some of its components were under investigation. To be specifically, the study investigated the relationships among strategic competence in the form of metacognitive strategy use reported by the Chinese EFL learners, task complexity perceived by these learners via their evaluation of the input format shown in the four computer-administered integrated speaking test tasks, and these learners' speaking performance indicated by their test scores.

1.6.2 Objectives

In light of the research scope, I assume the following four overarching objectives, which are addressed through answering the research questions (RQ):

a) To examine the retrospective data on the metacognitive strategies used by the Chinese EFL learners while they are performing the integrated speaking test tasks (see RQ1);
b) To examine the assumed task complexity of the integrated speaking test tasks induced by the cognitive load of the tasks (see RQ2 & RQ3);
c) To examine the Chinese EFL learners' speaking performance indicated by their test scores (see RQ4);
d) To examine the relationships among the metacognitive strategies (working independently and interactively) reported by the Chinese EFL learners, their speaking performance, and task complexity (see RQ5-RQ8).

1.6.3 Research questions

As delineated above, the four general research objectives are reframed as eight specific research questions concerning three research variables: Metacognitive strategy use, speaking performance, and task complexity. The research questions are listed as follows.

Research Question 1: What are the metacognitive strategies reported by the Chinese EFL learners in the context of the integrated speaking test?

Research Question 2: Is there any variability in the cognitive load on the Chinese EFL learners elicited by task complexity of the integrated speaking test tasks?

Research Question 3: What task characteristics affect task complexity of the integrated speaking test tasks perceived by the Chinese EFL learners?

Research Question 4: How do the Chinese EFL learners perform on each of the integrated speaking test tasks reflected by their test scores?

Research Question 5: What are the relationships between the metacognitive strategies reported by the Chinese EFL learners and their speaking performance in the context of the integrated speaking test?

Research Question 6: What are the relationships between task complexity and the Chinese EFL learners' speaking performance in the context of the integrated speaking test?

Research Question 7: What are the relationships between the metacognitive strategies reported by the Chinese EFL learners and task complexity in the context of the integrated speaking test?

Research Question 8: What are the relationships among the metacognitive strategies reported by the Chinese EFL learners, task complexity, and these learners' speaking performance in the context of the integrated speaking test?

1.7 Organisation of the Thesis

This thesis is organised into eight chapters. This chapter presents a brief introduction to this study in relation to the aims, the objectives, and the research questions. It also discusses the motivation, the research context, and the significance.

In Chapter Two, literature on the theoretical framework, including rationales for adopting Bachman and Palmer's (2010) framework of non-reciprocal language use, detailed descriptions of its component models, and the relevance of the framework to this study, is reviewed. A literature review on the research variables is included in Chapter Three. The two chapters lay a solid theoretical foundation for the study.

Chapter Four presents the research design and research methodology, which starts with the discussion on the overall research design, the specific design for the quantitative and the qualitative data, research sites, and sampling. It then describes and justifies the instrument development and validation, data collection procedures, and data analysis methods before ethical considerations are presented.

Chapter Five reports the quantitative results of the development and validation on the Metacognitive Strategy Inventory in Phase One. The reports aim to ensure that the newly developed questionnaire can be employed as a reliable self-report instrument in examining the Chinese EFL learners' metacognitive strategy use in the integrated speaking test. This chapter also provides the quantitative results from Phase Two to address Research Question 1, Research Question 2, Research Question 4, Research Question 5, Research Question 6, Research Question 7, and Research Question 8.

Chapter Six addresses both the student participants' and the teacher participants' perceptions of task complexity of the integrated speaking test tasks to answer Research Question 3. Furthermore, it depicts the students' in-depth views on their metacognitive strategy use in response to the integrated speaking test tasks. Results reported in this chapter complement those in Chapter Five for a comprehensive investigation of the research questions.

Chapter Seven entails a discussion on the main research results. Chapter Eight discusses the contributions, the implications and the limitations of this study. It also includes the suggestions for further research followed by my final thoughts as closing remarks. It should be noted that in line with previous studies of relevance (e.g., Bachman & Palmer, 2010; Davis, 2013; Purpura, 2016; Sato & McNamara, 2019), L2 in the thesis is used to refer to both the second language and the foreign language despite the differences

between the two concepts as defined by some researchers (e.g., Crystal, 2011; Qin, 2018; Richards & Schmidt, 2013).

Chapter Two

Literature Review: Part One

2.1 Chapter Overview

The literature review is divided into two parts: Part one regards literature on language assessment, which results in the adoption of Bachman and Palmer's (2010) framework of non-reciprocal language use as the theoretical framework; Part two primarily focuses on the three interdisciplinary variables of metacognitive strategies, task complexity, speaking, and relevant empirical studies.

This chapter presents the rationales for the theoretical framework. It begins with a review of language ability in modern language assessment followed by the review of the approaches and models germane to the theoretical framework. It also provides a detailed delineation of how the present study is underpinned by the framework.

2.2 Language Ability in Language Assessment

In modern language assessment, it is extensively recognised that the essence of language assessment is test-takers' language ability, and to understand language assessment is to understand language ability (e.g., Ellis et al., 2019; Hidri, 2018; Farhady, 2018; Read, 2016; Shoharmy, 2017; see also Bachman, 2005, 2007; Bachman & Palmer, 2010; Davis, 2013; McNamara, 2000; Weir, 2005).

2.2.1 Terminologies and definitions

Language ability is also labelled as language competence, language knowledge, language proficiency, communicative competence, to name a few (e.g., Bachman, 2005; Ellis et al, 2019; Farhady, 2018; McNamara, 2000; Purpura, 2008, 2016; Yi, 2012). It is difficult to be defined in that it cannot be observed directly in language assessment, which accounts for the ongoing debates on its definition even today (Ellis et al, 2019; Davis, 2014; Sato & McNamara, 2019). Traditionally, it was believed that language ability was composed of four skills: Listening, reading, speaking and writing. However, Bachman (1990) and Bachman and Palmer (1996, 2010) disagree with this view. Theoretically inspired by Canale and Swain's (1980) seminal work on communicative

competence, which is influenced by Hymes' (1972) language use theory, Bachman (1990) and Bachman and Palmer (1996, 2010) argued that when we define language ability, we need to take into each specific assessment situation into consideration. For this purpose, language ability should be defined as a construct and this construct is a specific definition of language ability that provides the basis for a specific assessment task and for the interpretation of the scores derived from the task. In a more explicit and precise manner, Chapelle et al. (2011) explained that test-takers' language ability indicates their ability to integrate language knowledge (e.g., grammar and vocabulary) with strategies required for accomplishing a particular goal.

In fact, in modern language assessment, the debate on language ability often results in a specific model or a framework (Bachman, 2007, Purpura, 2016). Some researchers (e.g., Fulcher, 2014; Fulcher & Davidson, 2007) consider models and frameworks as different concepts: Models refer to theoretical interpretations of the meaning communicated in a second language, whilst frameworks indicate selecting skills and abilities from a model according to a given assessment context. On the other hand, some researchers (e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; Farhady, 2018; Purpura, 2016; Zhang, 2017) use the two terms interchangeably in defining language ability. An example of such a practice is Lado's (1961) ideas on language ability. Purpura (2016) termed it as a skills-and-elements model of L2 proficiency, whereas Farhardy (2005) referred to it as a skill framework. With reference to such interchangeability, I used "model" and "framework" interchangeably.

Among all the models developed to define language ability, some of them have profound influence on the evolution of this construct. For instance, the above mentioned Lado's (1961) model in which language ability is assumed to be the sum of individual language components such as phonology, structure and lexis or skills like reading, listening, speaking or writing (Purpura, 2008, 2016); Oller's (1983) hypothesis of the unitary competence where language ability is a unitary factor in various forms of language components and skills; Canale and Swain's (1980) Communicative Competence Model, Bachman's (1990) Communicative Language Ability Model, and Bachman and Palmer's (2010) Language Ability Model, all of which highlight test-takers' language ability for language use.

Though these models are different from one another, they demonstrate one commonality: “language ability is a complex construct with multiple dimensions” (Gu, 2014, p. 112). Among the debating voices on the definitions, Farhady (2018) commented that the unavailability of a solid definition of language ability is due to the lack of a theory for defining the construct. In reviewing language assessment, Purpura (2016) explored the reason behind the unavailability, positing that defining language ability is “probably the most compelling and enduring challenge in L2 assessment” (p. 193). He also pointed out that to meet the challenge, researchers in this area had taken various approaches to define language ability.

2.2.2 Approaches to defining language ability

According to Purpura (2016), approaches adopted by researchers to define language ability are categorised as the trait-based approach, the task-centred approach, the interactionist approach and the socio-interactional approach, which correspond to Bachman’s (2007) interactional or interactive ability approach, the real-life approach, and the internationalist approach. Language ability defined in the interactive ability approach is measured as a mental activity and it is conceptualised as an individual’s capacity to use language. Bachman (1990, 2007) contended that this approach should incorporate two frameworks: One has to do with language ability and the other is about test methods. The internactionalist approach was originally proposed by Chapelle (1999) from the social interactional perspective. It explains test-takers’ performance consistency from the aspects of test-takers’ traits, the features of the contexts in which tests are performed, and the interactions between them (Bachman, 2007; Purpura, 2016). The real-life approach is also termed as “task-based performance assessment” approach which focuses on to what degree a test can replicate the real-world non-testing language use situations and whether test-takers’ performance can predict their future performance in a real-world language use settings (Bachman, 1990; 2007; Ellis et al, 2019; Purpura, 2016). In comparing the three approaches, Bachman (2007) pointed out that the interactionalist approach lacks theoretical evidence associated with language assessment, and the real-life approach does not accommodate the interactional properties of language assessment. In a different vein, the interactive ability approach is more systematically grounded, approximating the process of language assessment: It

encompasses not only language user parameters, language use task characteristics but the interactions involved as well.

2.2.3 Language ability models in the interactive ability approach

As Skehan (1998) interpreted that the interactive ability approach is to develop a model of abilities underlying language assessment, the construct of language ability is typically conceptualised within a specific model or a framework as delineated earlier (Bachman, 2007). However, among these models only three are proposed in the interactive ability approach: Bachman's (1990) Communicative Language Ability Model, Bachman and Palmer's (1996) Language Ability Model, and Bachman and Palmer's (2010) revised Language Ability Model with the latter two models being the refined version of Bachman's (1990) model (Bachman, 2007; Purpura, 2016). It is obvious that all the three models pertain to Bachman, and it explains why Skehan (2018) labels the three models as Bachman's approach which he describes as "well established, influential and clear" (p. 291).

2.2.3.1 Bachman's (1990) Communicative Language Ability Model

By synthesising prior literature on language ability, Bachman (1990) developed his Communicative Language Ability Model based on Canale and Swain's (1980) Communicative Competence Model. In Canale and Swain's (1980) multi-componential model, communicative competence is a synthesis of knowledge and skills needed for communications, and it is divided into three categories: Grammatical competence, sociolinguistic competence and strategic competence. Later, Canale (1983) updated this model and added discourse competence (Bagarić, & Djigunović, 2007; Purpura, 2008). Strategic competence in this model plays a compensatory role in language use, through which language users can solve their communication problems. Its capacity is exemplified by the use of discourse markers to flow conversations for communicative purposes.

In a different vein, Bachman (1990) regards strategic competence as language users' mental ability to implement language competence for communications in language assessment. He places strategic competence "at the heart of all normal communications" in his Communicative Language Ability Model (Ellis et al., 2019, p. 244), and gives it

a central role operating in three areas: Goal setting; assessment; and planning (Skehan, 2018; Ślęzak-Świat, 2008).

In addition to strategic competence, the Communicative Language Ability Model also incorporates language competence and psychophysiological mechanisms. Bachman interprets language competence as specific knowledge in the form of language used in communications. Psychophysiological mechanisms are the neurological and psychological processes involved in language use such as sound and light. The three components of the model interact with language use contexts and language users' structural knowledge via strategic competence. Later, the Communicative Language Ability Model was reconfigured by Bachman and Palmer (1996, 2010) as the Language Ability Model in which the crucial role of strategic competence in one's language ability is more conspicuously highlighted.

2.2.3.2 Bachman and Palmer's (1996, 2010) Language Ability Model

In the Language Ability model revised in 1996, language ability is part of individuals' characteristics which also include non-cognitive factors (topic knowledge, personal characteristics and affective schemata), and these individual attributes interact with one another, affecting test performance. Strategic competence in this model is viewed as the metacognitive strategies of goal setting, assessment and planning as it is in Bachman's (1990) Communicative Competence Model. More recently, Bachman and Palmer (2010) further revised the model so that strategic competence is explicitly conceived as a set of higher order metacognitive strategies that "provide a management function in language use, as well as in other cognitive activities" (Bachman & Palmer, 2010, p. 48).

Compared with the model proposed in 1996, strategic competence in the newly revised model subsumes different metacognitive strategies: Goal setting, appraising and planning. Goal setting concerns language users' decision on what they are going to do for a given language use or language assessment task. It helps language users to identify, to choose and to decide their intended language tasks. Appraising helps language users to appraise the language tasks intended and the possibility and feasibility of completing the tasks. In addition, appraising helps language users appraise their own knowledge for the successful completion of the intended tasks. The third component of the strategic competence is planning which is about how language users decide how to use what they

have for completing the tasks intended. Planning, thus, involves the selection of one's topic knowledge and language knowledge, designing plans to use this knowledge and choosing one specific plan in reaction to the tasks given or intended. Table 2.1 shows the latest refined model of the strategic competence.

Table 2.1 *Bachman and Palmer's (2010) Strategic Competence Model*

Strategies	**Definitions**
Goal setting	Identifying the intended tasks
	Selecting tasks
	Deciding whether or not to complete the selected tasks
Appraising	Appraising task characteristics to determine the possibility of task completion and the relevant resources needed
	Examining the prior knowledge available
	Evaluating task performance
Planning	Selecting one's prior knowledge available for task completion
	Formulating plans to complete the tasks
	Selecting one particular plan for task completion

Note. This table is adapted from *Language Assessment in Practice* by L. Bachman and A. Palmer, 2010, p. 49. Copyright 2010 by Oxford University Press.

Although Bachman and Palmer (1996, 2010) consider language ability as part of individual attributes, they separate it from other individual attributes because in language use situations, individual attributes are "more peripheral attributes of language users" (p. 43) than language ability. On this basis, Bachman and Palmer argue that test-takers' characteristics in language use should be composed of two traits: Language ability and attributes of individuals with language ability situating in the centre. In other words, strategic competence in the newly revised model is categorised as the focal individual attributes in contrast with the peripheral features of individuals which include topic knowledge, affective schemata, personal attributes and cognitive strategies (Zhang, 2017). Bachman and Palmer also believe that test-takers' strategic competence is in charge of the utilisation of individuals' attributes such as their language knowledge and topic knowledge for interactive responses to language use domains, and it serves as

a mediator between language users and language use tasks, impacting the language users' performance.

Another component of the Language Ability Model (1996, 2010) is language knowledge in contrast with the language competence in Bachman's (1990) Communicative Competence Model. Language knowledge in the two versions of Language Ability Model is constituted by organisational knowledge and pragmatic knowledge. Organisational knowledge accounts for how individuals' utterance and texts are organised for a given task of language use or language assessment. By contrast, pragmatic knowledge is responsible for how individuals' utterance and texts are associated with their communicative goals and with the properties of language use settings. Figure 2.1 shows the newly revised Language Ability Model.

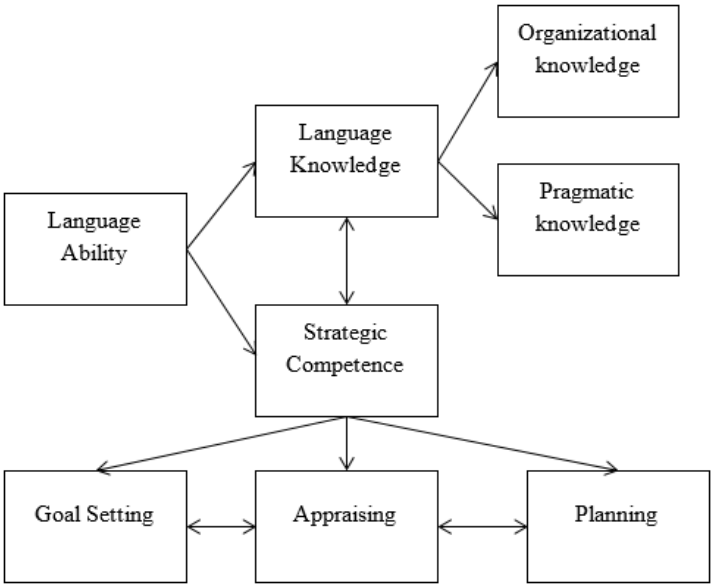

Figure 2.1 *Language Ability Model*

Note. This Model is adapted from *Language Assessment in Practice* by L. Bachman and A. Palmer, 2010, pp. 44-49. Copyright 2010 by Oxford University Press.

Of the three models, Bachman's (1990) Communicative Language Ability Model is recognised as "one of the most comprehensive models of L2 proficiency" (Purpura, 2016, p. 195), and "the most influential multi-componential model of communicative language ability" (O'Sullivan, 2011, p. 13). By the same token, Bachman and Palmer's

(1996) Language Ability Model is commented as “one of the most influential models to date” (Cox, 2020), exerting profound influence on L2 assessment (McNamara, 1996, Skehan, 1998; O’Sullivan, 2000; Fulcher, 2014).

2.2.4 Reconceptualising language ability

As seen above, the language ability models drawn upon by Bachman (1990), and by Bachman and Palmer (1996, 2010) in the interactive ability approach portray the interactional process of language assessment. Nevertheless, these models are not without drawbacks. Strategic competence in Bachman’s (1990) Communicative Language Ability Model only refers to test-takers’ mental ability to implement language competence for the purpose of communications in language use. Due to this, it is not applicable in language use situations where no inter-individual communications take place such as the increasingly popular computer-assisted tests (Chapelle & Voss, 2016; Luoma, 2004). Likewise, Bachman and Palmer’s (1996) model has been criticised for some limitations including the exclusion of cognitive strategies, an important variable in language assessment (Skehan, 1996; Ślęzak-Świat, 2008). Though in their (2010) newly refined Language Ability Model, language ability is defined from a more comprehensive perspective to reflect the language assessment process, the independent application of all the three models per se in examining language assessment essentially focalises on test-takers. Henceforth, they cannot be applied in the investigation of the interactions involved in language assessment between test-takers and test tasks. Consequently, other models or frameworks on test tasks have to be used in parallel to indicate the interactions (Luoma, 2004).

More broadly, in language assessment, a consensus has been reached among some researchers (e.g., Bachman & Palmer, 1996; 2010; Hidri, 2018; Weir, 2005) that in defining language ability, the context of language use should be taken into account. They believe that individuals’ use of the specific linguistic knowledge such grammar and vocabulary and their use of strategies depend on the context where language performance takes place. The limitations of the three interactive models and the consensus on conceptualising language ability in a language use context justify Bachman and Palmer’s (1996, 2010) ever-increasing emphasis on conceptualising language ability in an interactional framework of language use. In their own words,

"Language ability must be considered within an interactional language use framework" (Bachman & Palmer, 1996, p. 62); "We also need to define language ability in a way that is appropriate for each particular assessment situation" (Bachman & Palmer, 2010, p. 43).

Although test-takers' language ability or a particular aspect of their language ability is of the primary interest of language assessment, characteristics of test tasks are equally critical. This salience is because the correspondence of the test task characteristics to the non-assessment situations or the real language use situations will determine to what extent the oral interpretation on test-takers' language ability is meaningful. In other words, whether a test-taker's performance in a language test can be generalised to his or her language ability in a wider real-life language use setting will depend on the authenticity of test tasks (Bachman & Palmer, 1996, 2010).

To achieve such authenticity, Bachman and Palmer (1996, 2010) proposed corresponding language test performance to language use situations. They contended that language assessment assesses test-takers' language ability in accordance with their performance on a specific test in order to generalise individual's ability to handle similar but real-world language use situations. If, as expected, individuals' performance can help to predict or evaluate their real capacity to use language appropriately in a real and future language use setting, the correspondence between language test performance and language use should be established. Bachman and Palmer further proposed that the establishment could be achieved with a conceptual language use framework in which test-takers' performance is interpreted as a particular example of language use. They believe that such framework allows language testers to refer to standardised templates in describing the essential properties of both test-takers' performance on language tests and their performances in non-test language use situations as well (Bachman, 2007).
Bachman and Palmer's (1996, 2010) emphasis on situating language ability in a language use framework is also demonstrated in their argument for the intended usefulness of language assessment through the framework of Assessment Use Argument (AUA) (Bachman, 2005; Bachman & Palmer, 2010). In explaining the AUA, Bachman and Palmer stated that the meaningful definitions and interpretations of language ability construct should be based on real language use situations. Alternatively stated, language

ability should be language-use dependable. For instance, if the assessment is used for diagnostic purpose in classroom instructions, the definition of language ability will depend on a language instructional syllabus; if the use of assessment is for the target language use domains (a specific setting outside of the test itself that requires the test-takers to perform language use tasks) such as academic admission or employment decision, the construct will be defined according to a need analysis. Though the AUA aims at the intended usefulness of language assessment, it does underscore the reliance of language ability on language use, which supports Bachman and Palmer's proposal of conceptualising test-takers' language ability in a global language use framework.

In fact, the obvious changes in Bachman and Palmer's wording in defining language ability and test tasks can be understood as additional support for their argument on conceptulasing language ability in a language use framework. In his book published in 1990, Bachman used "communicative language ability" and "test methods facets" to refer to language ability and test tasks, respectively. However, in his cooperation with Palmer in 1996, the two authors situated the two concepts in the context of language use in their book. The chapter on language ability is titled "Describing language ability: Language use in language tests" (p. 60). Similarly, they named the chapter on test tasks as "Describing tasks: Language use in language tests (p. 43). The discrepancies in their wording from that of Bachman (1990) imply that they began to interpret and promote language assessment within the context of language use. The wording in their more recent book continued to push forward such an effort: Bachman and Palmer (2010) juxtaposed language use with language ability and test tasks purposefully by naming the chapters on language ability and test tasks as "Describing language use and language ability" (p. 33) and "Describing characteristics of language use and language assessment tasks" (p. 59). Such constant and explicit modifications in wording not only reveal Bachman and Palmer's refinement of their understanding of language assessment, but also go some way to rationalise their repeatedly-emphasised proposal: The examination of language ability should be conducted in an interactional framework of language use.

2.3 Interactional Frameworks of Language Use

According to Bachman and Palmer (1996, 2010), language use refers to the activity of creating or interpreting individuals' intended meanings with discourses in response to a

specific situation. The intended meanings in interactive negotiations between two or even more individuals are also defined as language use. Typically, language use involves internal and external interactions. Internal interactive language use refers to the interactions among individual language users' attributes (e.g., age, sex and nationality), topic knowledge, affective schemata and cognitive strategies as delineated previously. External interactions imply the interactions between language users. The characteristics of language use situations include the language that the language users are processing in either written or spoken forms, and the other language users. With reference to the definitions, Bachman and Palmer posited that there are two types of language use: Non-reciprocal language use in which only one language user is engaged; and reciprocal language use which involves two or more language users.

Grounded in their definitions of language use, Bachman and Palmer (1996, 2010) proposed two visual frameworks: The simpler framework of non-reciprocal language use and the more complicated framework of reciprocal language use. Bachman and Palmer noted that within the two conceptual language use frameworks, to illustrate the interactions between individuals and language use tasks in language use situations, two set of characteristics should be included: One set regarding test-takers or language users and the other one concerning test tasks or language use tasks. The set of characteristics of test-takers is presented in their (1996, 2010) Language Ability Model and the Model of Individual Attributes (see Section 2.4), and the characteristics of test tasks are shown in their (1996, 2010) Framework of Task Characteristics and Framework of Language Use Task Characteristics. As discussed above, the revised version of Language Ability Model (2010), compared with the earlier one (1996), include the cognitive strategies, and hence provides a closer replication of language ability. In a different vein, except for their names, there are no fundamental difference between the two versions of the test task framework,

In comparison with the Language Ability Model (2010), Language Tasks Characteristics Model (2010) is more complex, as it is comprised of five components: Setting, rubrics, input, response and the relationship between input and response. The characteristic of setting is comprised by physical characteristics, participants and time of a specific language use or language testing task. Rubrics is characterised by

instructions such as the language of instructions, the channel through which the instructions are given, and the specifications of structures and procedures for test-takers to follow to finish a given task. Time allotment and recording methods are the other two properties of rubrics. In terms of input and the expected response, their features are demonstrated by the format and the language used. According to Bachman and Palmer (2010), input constitutes the material in a test task that test-takers are expected to process and to which they are expected to respond. It is regarded as the problem that test-takers have to solve, and test-takers' ability to process the input will affect their performance. Finally, test task characteristics are illustrated by the relationship between the input and the expected response which is constituted by the type of external interactions among test-takers, and between test-takers and the equipment and materials used for the language use tasks. Such a relationship is also characterised by the scope of the relationship (whether it is narrow or broad), and the directness of the relationship (whether it is direct or indirect). Figure 2.2 shows the Language Use Tasks Characteristics Model (2010), and with reference to the model, it is apparent that the integrated speaking test tasks under investigation demonstrate the task characteristic of input as noted in 1.3.

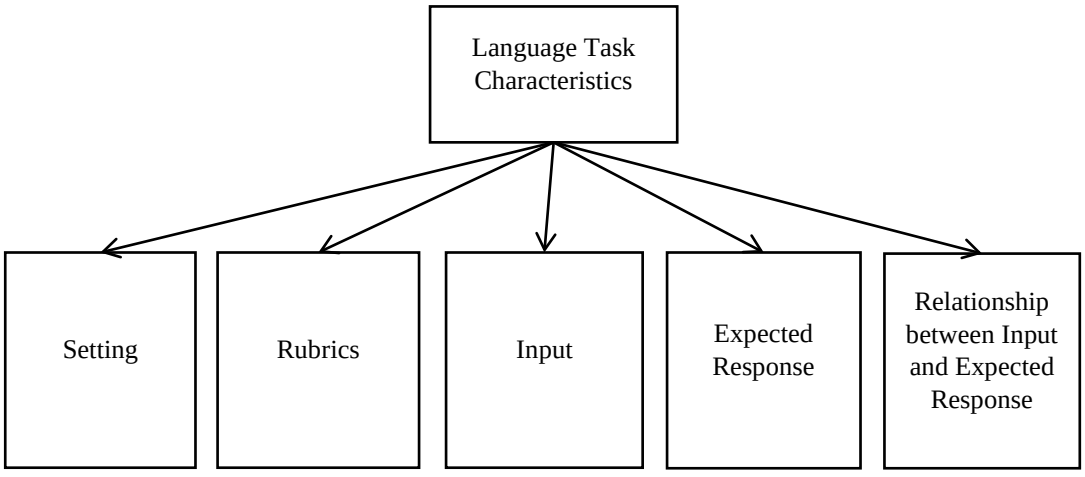

Figure 2.2 *Language Use Tasks Characteristics Model*

Note. This Model is adapted from *Language Assessment in Practice* by L. Bachman and A. Palmer, 2010, pp. 66-68. Copyright 2010 by Oxford University Press.

2.4 Theoretical Framework

Indeed, the differences illustrated by the models on language ability and the changes in the wording on task characteristics discussed above differentiate the frameworks of language use developed by Bachman and Palmer in 1996 and 2010 respectively. It is therefore apparent that the newly refined interactional frameworks of language use are closer to the language use situations in the real world. This closer correspondence explains why I framed my study in a new framework. Moreover, in the new frameworks, the framework of reciprocal language use focalises on individual-to-individual interactions, whilst the framework of non-reciprocal language use illustrates the interactions within test-takers' attributes and between these attributes and test task characteristics in the language use settings in which there are no individual-to-individual conversations. Such properties establish the compatibility between this non-reciprocal framework and the present study, which rationalises why this framework, a refined version of the framework (1996) that "has been very influential in defining the construct of language ability" (Hidri, 2018, p. 3), serves as the theoretical framework of the study. Figure 2.3 displays the framework of non-reciprocal language use.

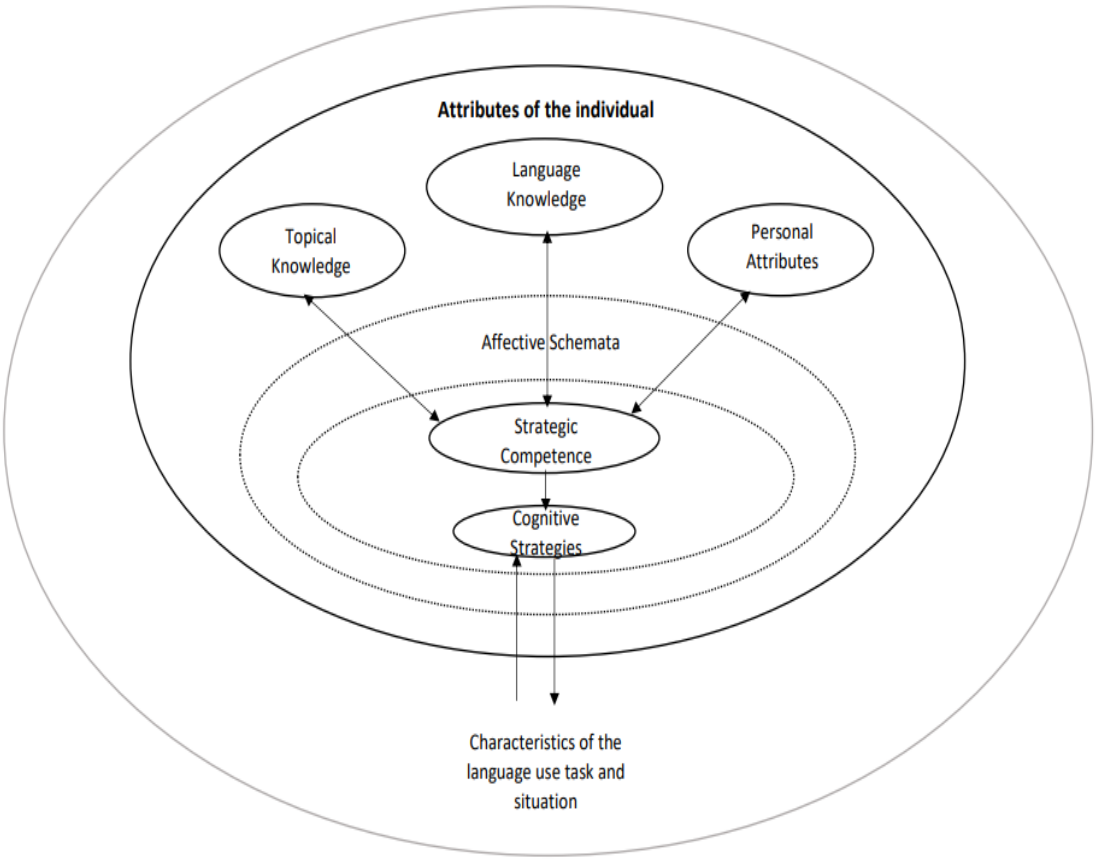

Figure 2.3 *Non-reciprocal Language Use Framework*

Source: Language Assessment in Practice" by L. Bachman and A. Palmer, 2010, p. 36. Copyright 2010 by Oxford University Press.

As the above review shows, the framework of non-reciprocal language use accommodates the Language Ability Model and the Language Use Task Characteristics Model to duplicate the two sets of the characteristics regarding the test-takers and the test tasks respectively. In addition, because Bachman and Palmer purposefully separated language ability from individual attributes to highlight its core role in language use, the framework also includes the Model of Attributes of Individuals. This model is composed of personal properties including age and sex, topic knowledge/content knowledge, knowledge schemata/real world knowledge, affective schemata which refers to individuals' feelings associated with one's topical knowledge for evaluating the

characteristics of language use tasks and cognitive strategies. Figure 2.4 shows the Model of Individual Attributes.

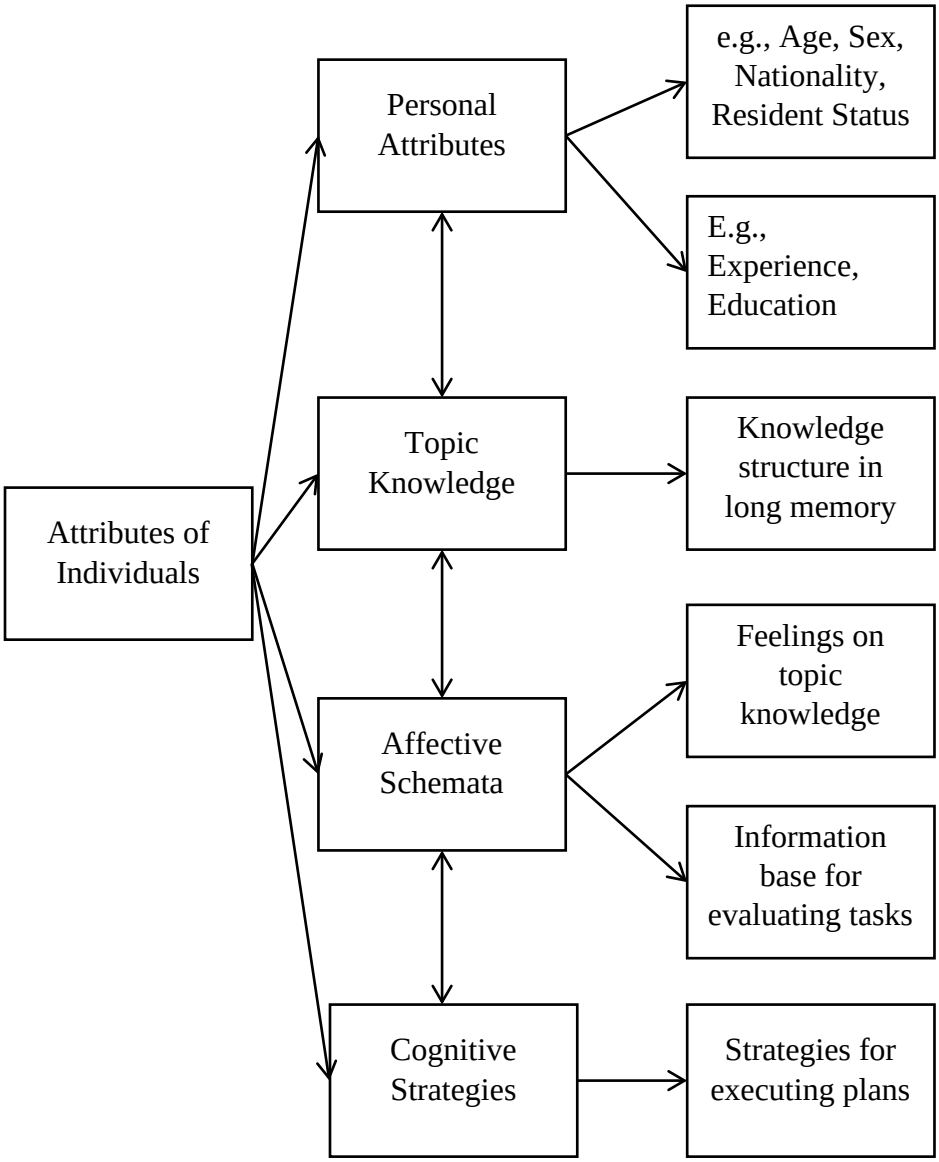

Figure 2.4 *Model of Individual Attributes*

Note. This model is adapted from *Language Assessment in Practice* by L. Bachman and A. Palmer, 2010, pp. 41-43. Copyright 2010 by Oxford University Press.

2.5 Relevance of the theoretical framework to this study

From Figure 2.3, it can be seen that the non-reciprocal framework includes many subcomponents. However, only two of them were involved in this study: The strategic

competence related to test-takers and the task characteristic of input reflected in the integrated speaking test tasks which indicates varying degrees of task complexity.

2.5.1 Strategic competence model

In L2 assessment, because of the extensive influence of Bachman and Palmer's (1996, 2010) models on language ability, strategic competence has been recognised as test-takers' metacognitive strategies operating in their test performance. Hence, it is quite common that researchers treat an individual's strategic competence as his or her use of metacognitive strategies in empirical studies (Huang, 2013; Seong, 2014). Examples include Barkaoui et al. (2013), Fernandez (2018) Huang (2013), Phakiti (2003, 2008, 2016), Purpura (1997, 1998; 1999, 2008; 2013), Song (2005), Swain et al. (2009), Youn, and Bi (2019), Zhang (2017), and Zhang, Goh and Kunnan (2014). Nevertheless, with regard to what metacognitive strategies are actually used by the test-takers in the real language tests, answers from these researchers are inconclusive and such a result is due to the ambiguity plaguing in the conceptualisation of strategic competence as discussed earlier and the insufficiency of empirical studies for validating Bachman and Palmer's (2010) strategic competence model (Ellis et al., 2019; Huang, 2013; Phakiti 2003, 2008, 2016; Purpura 1998; 1999; 2008, 2013; Seong, 2014). Consequently, although metacognitive strategies are conceived as goal setting, planning and appraising in the model, researchers (e.g., Barkaoui et al., 2013; Huang, 2013; Fernandez; 2018; Phakiti 2003, 2008, 2016; Purpura, 2008, 2013; Swain, 2009) commonly take an exploratory approach to investigate strategic competence in line with the literature on L2 assessment, metacognition, and learning strategies. Henceforth, strategic competence operated as various forms of metacognitive strategies in these empirical studies. For example, Barkaoui and colleagues (2013) discovered that the metacognitive strategies used by the Chinese participants were identifying the purpose of the task, setting goals, evaluating previous performance, and evaluating content of what heard/said. By contrast, in Zhang's (2017) study, metacognitive strategies identified were assessing the situation, monitoring, self-evaluation and self-testing. Given the actual situations where strategic competence is examined in L2 assessment, I am aware of the lack of consensus about what the actual metacognitive strategies used by test-takers are. Therefore, I also took an exploratory approach from an interdisciplinary perspective to examine the Chinese EFL learners' metacognitive strategy use within the theoretical framework.

2.5.2 Input

In the theoretical framework, the integrated speaking test tasks reflect the task characteristic of input which is further broken down into two factors: The input format and the input language (Bachman & Palmer, 2010). Since characteristics of language such as the vocabulary and grammar are not of the research interest (see Section 1.3.1), I did not explain input language in this thesis. On the other hand, the format of input corresponds to the task characteristics of the integrated speaking test tasks, and hence is reviewed succinctly below.

2.5.2.1 Format of input

The format of input is comprised of seven characteristics: (a) Channel of input which can be aural, visual or both, such as a video of lecture; (b) form of input based on language or non-language, such as pictures, gestures and actions; (c) language of input which can be test-takers' native language or the target language; (d) length or time of input which refers to the amount of materials that test-takers process and the time that is allotted for them to process the test task; (e) vehicle of input such as audio or visual; (f) degree of speediness related to the rate at which the input information is presented to test-takers; and (g) type of input including item of input, prompt of input, and input for interpretation. An item of input consists of the input information used to elicit responses from test-takers, a prompt of input associated with the input in the form of a directive for eliciting test-takers' expected responses, and the input for interpretation which contains written or oral language presented to test-takers for the completion of a test task.

2.5.2.2 Integrated speaking test tasks

Integrated speaking test tasks integrate language use activities to make authentic language use in testing contexts. Based on the working model of language use in an authentic academic context, integrated speaking tests are theoretically considered as an expanded version of the Model of Communicative Language Ability (Bachman, 1990; Brooks & Swain, 2014; Canale & Swain, 1980). They are developed to replicate a real-like language use task so that test-takers' ability to communicate in English can be measured (Bridgeman et al., 2012). It is believed that if test-takers do well on these tests,

they have shown that they possess the abilities required in real language use situations where multiple language skills are needed (Luoma, 2004).

Because of such authenticity, integrated speaking test tasks are used in many popular high-stakes tests such as TOEFL speaking section (Hughes & Reed, 2017). The four integrated speaking test tasks in this study are borrowed from the test. Task 3 and Task 4 are reading-listening-speaking tasks with 30 seconds for preparations and 60 seconds for speaking, while Task 5 and Task 6 are listening-speaking test tasks with 20 seconds for preparations and one minute for speaking. Furthermore, Task 3 and Task 5 are on campus life situations whilst Task 4 and Task 6 relate to academic lectures. In terms of task type, Task 3 requires test-takers to give an oral summary of the speaker's opinion whilst Task 4 and Task 6 ask test-takers to use the examples given to illustrate an academic concept presented by the speaker. In contrast, Task 5 is about providing solutions to solve a specific problem given (e.g., Alderson, 2009; Crossley & Kim, 2019; Huang et al., 2018; Hughes, 2017).

In light of the input format, it is obvious that the four speaking test tasks illustrate the variability in the four characteristics of the input format: Topic characteristics or the item of input, input for interpretation/length, time for preparations and prompt (Barkaoui et al., 2013). Table 2.2 displays the variability in input, which essentially indicates task complexity to varying degrees in the four integrated speaking test tasks.

Table 2.2 *Variability in Input Format in the Speaking Test Tasks*

Tasks	Topic Characteristics	Input for interpretation/Length	Time (Preparations)	Prompt
Task3	Campus-life Situation	R-L-S	30s	Summary
Task4	Academic Lecture	R-L-S	30s	Summary
Task5	Campus-life Situation	L-S	20s	Problem-solving
Task6	Academic Lecture	L-S	20s	Concept-illustrating

Note. R = reading, L = listening, S = speaking, PK = prior knowledge.

To summarise, the delineation of the relevance of the strategic model and of the task characteristic of input to this study reveals how the study is underpinned by Bachman and Palmer's (2010) framework of non-reciprocal language use as illustrated in Figure 2.5.

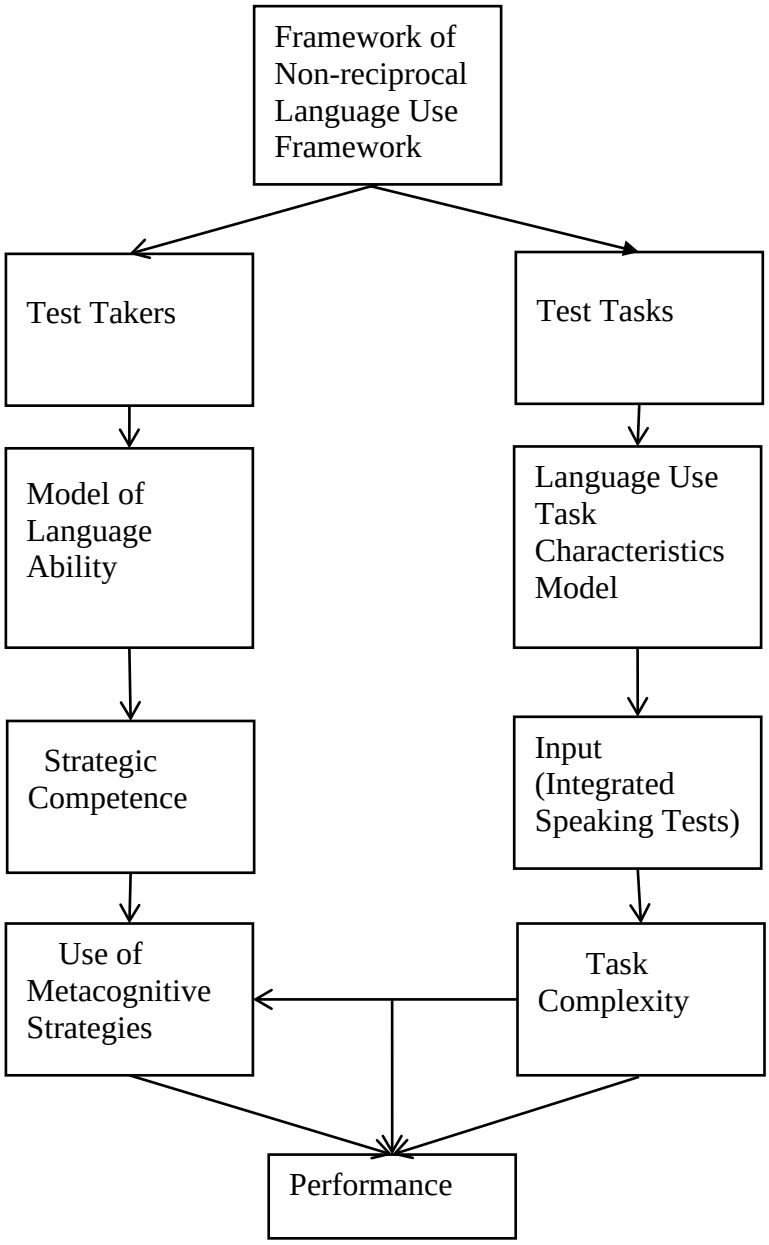

Figure 2.5 *This Study Framed in the Theoretical Framework*

Chapter Three

Literature Review: Part Two

3.1 Chapter Overview

This chapter presents a review regarding metacognitive strategies, task complexity and speaking performance in the arenas of language assessment, second language acquisition and educational psychology. Underpinned by the theoretical framework reviewed in Chapter Two, Chinese EFL learners' strategic competence is defined as their ability to use metacognitive strategies, and it is examined in an exploratory approach from the interdisciplinary perspective (Bachman, 1990; Bachman & Palmer, 1996, 2010). As such, the review of the literature on strategic competence is essentially a comprehensive literature review of metacognitive strategies. The chapter also rationalises the examination of task complexity reflected in the integrated speaking test tasks within Robinson's (2015) Triadic Componential Framework. In addition, the review of speaking performance is framed in Kormos' (2011) Bilingual Speech Production Model.

3.2 Metacognitive Strategies

As reviewed in Section 2.5.2.1, in an actual process of test performance, metacognitive strategies work in diverse forms, and researchers in L2 assessment commonly examine this construct from perspectives of metacognition and language learning strategies. In line with this common practice, to formulate the working definition of metacognitive strategies under investigation, my literature review on this construct was also conducted from the two above perspectives in relation to L2 assessment.

3.2.1 Metacognitive strategies in metacognition

Metacogitive strategies, originated from the field of psychology, especially in relation to how metacognition is understood, are generally regarded a pivotal metacognitive element. According to McNamara (1996), the examination of any strategic competence model should be cross-referenced with the literature on metacognition. My review of this construct, therefore, started with the relevant literature born in the research field of metacognition.

3.2.1.1 Definitions and taxonomies of metacognition

Metacognition was introduced by Flavell (1976) who postulated that "metacognition refers to one's knowledge concerning one's own cognitive process and products or anything related to them" (p. 32). Since its introduction, metacognition has triggered debate about its definition and components (e.g., Qin. 2018; Zhang & Qin, 2018; D. Zhang & Zhang, 2019; L. Zhang & Zhang, 2018). Despite this, it is acknowledged that the foundational research on metacognition takes root in two frameworks proposed by Flavell (1979) and Brown (1987) (Nazarieh, 2016; Sperling et al., 2012).

Flavell's (1979) Framework of Metacognition

Flavell's Framework of Metacognition is also known as Model of Cognitive Monitoring (Nazarieh, 2016; Tarricone, 2011) which includes four elements: Metacognitive knowledge, metacognitive experience, tasks or goals, and strategies or actions. Metacognitive knowledge is the knowledge of individuals, tasks or goals, and strategies or actions which act either independently or interactively to exert impact on the course and the outcome of cognitive activities. Metacognitive experiences or metacognitive regulation are consciously cognitive and affective, and are associated with learning (Livingston, 2003; Mahdavi, 2014). Tasks or goals are an individual's cognitive objectives. The strategies or actions are cognitive activities and or other behaviours that the individual uses to achieve the objectives. Flavell further divided metacognitive knowledge into three subcategories on persons, tasks and strategies. Personal knowledge is about how individuals learn and how personal factors affect their learning activity. Task knowledge is about task demand. Strategy knowledge refers to the knowledge of strategies including knowledge of cognitive strategies, metacognitive strategies and conditional knowledge of when and where to use these strategies (Anderson, 2012; Livingston, 2003;Tarricone, 2011; D. Zhang & Zhang, 2019; L. Zhang & Zhang, 2018).).

Put simply, metacognition empowers learners to use strategy A instead of B in the task X as contrasted in the task B when addressing a particular task for a learning goal (Flavel, 1979). Figure 3.1 illustrates Flavell's Framework of Metacognition.

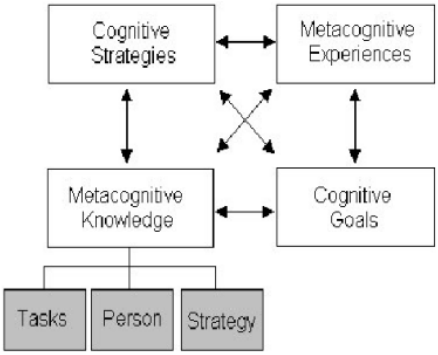

Figure 3.1 *Flavell's (1979) Framework of Metacognition*

Source: An overview: Metacognition in education, by M. Mahdavi, 2014, *International Journal of Multidisciplinary and Current Research*, 2 (6), p. 530. Copyright 2014 by IJMCR.

Brown's (1987) framework of Metacognition

In Brown's framework, metacognition is parsed into two components: Knowledge of cognition and regulation of cognition. Knowledge of cognition is the knowledge of one's cognitive activities, which includes applying thoughts and ideas about cognitive activities of oneself or other people. Regulation of cognition indicates planning, monitoring and evaluating a learning or problem-solving activity.

Knowledge of cognition is further divided into three subcomponents: Declarative knowledge, procedural knowledge and conditional knowledge. Declarative knowledge contains knowledge of oneself, task and strategy. It is about understanding oneself as a learner and the factors that influence one's performance. Procedural knowledge, represented as strategies, concerns performing tasks. A learner with a high degree of procedural knowledge performs tasks in a more automatic manner, hence is more likely to own a larger repertoire of strategies at disposal. Learners assisted by procedural knowledge can sequence strategies effectively as they know how to use strategies for solving problems and for performing tasks. Conditional knowledge involves the knowledge of using declarative and procedural knowledge according to the real conditions. Conditional knowledge allows learners to use their strategies effectively and efficiently by allocating resource selectively (e.g., Qin, 2018; Teng, 2016; Zhang & Qin, 2018; Zhang & Zhang, 2018).

Regulation of cognition or metacognitive regulation is also referred to as metacognitive control (Nazarieh, 2016), metacognitive experience, and metacognitive strategies (Brown, 1987; Livingston, 2003). It involves actions that individuals take to control their cognitive activities for reaching a goal, including planning, and monitoring the individuals' cognitive activities and evaluating the results of these cognitive activities (Nazarieh, 2016; Livingston, 2003, Papaleontiou-Louca , 2003, 2008). Brown (1987) proposed that planning helps individuals decide how to finish a learning task; monitoring checks their learning process and plans for the learning task. Evaluating is associated with the outcome of the individuals' efforts in processing the learning task or with the results of a plan for improving their future learning activities.

Brown (1987) suggested that metacognitive regulation is unstable and more task-dependent than metacognitive knowledge that is stable and age-dependent. Such characteristics are well interpreted by Nazarieh (2016): "Adults might not use strategies when solving a simple problem (unstable); young learners might not have the ability to monitor and regulate their strategies (age independent)" (p. 63). Figure 3.2 shows Brown's Framework of Metacognition.

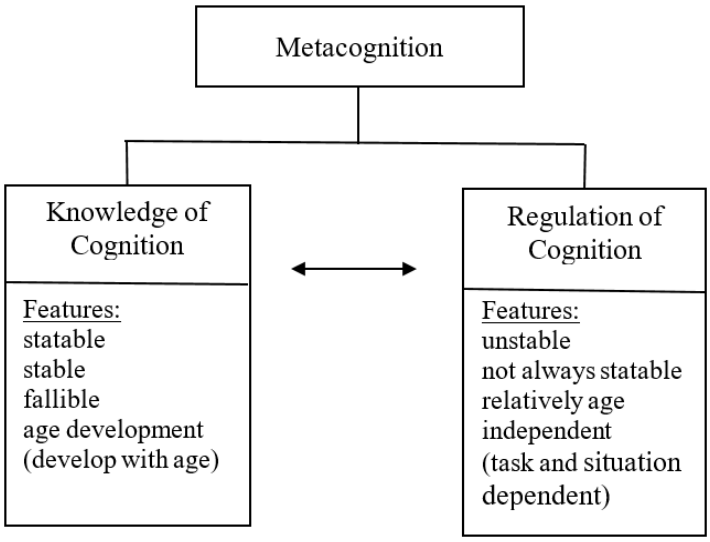

Figure 3.2 *Brown's (1987) Framework of Metacognition*

Souce: Conceptualising and Assessing Metacognitive Development in Young Children by L.M. Marulis, 2014, p. 5. Copyright 2014 by L. M. Marulis.

3.2.1.2 Metacognitive strategies as a component of metacognition

Essentially, both Flavell's model (1979) and Brown' framework (1987) reveals the influential role of metacognitive strategies, albeit in different terms and definitions. In Flavell's model, both metacognitive knowledge and metacognitive experience are closely related to metacognitive strategies: One subcategory of metacognitive knowledge is knowledge of metacognitive strategies, and metacognitive experience involves the use of metacognitive strategies. Likewise, in Brown's framework, the instability of metacognitive regulation and its dependence to tasks indicates its role as metacognitive strategies (Nazarieh, 2016; Livingston, 2003) because O'Malley and Chamot (1990) pointed out that metacognitive strategies might not be stable and be more task-dependent in comparison with metacognitive knowledge. Some researchers also argue that Flavell's and Brown's views on the constitution of metacognition can be generally categorised into two components: Metacognitive knowledge and metacognitive regulation or metacognitive strategies (Livingston, 2003; Tarricone, 2011). Oxford (2011) even explicitly contended that Flavell's mode of metacognition can be understood as metacognitive knowledge plus the use of metacognitive strategies, as she interpreted metacognitive regulation as being equal to the use of metacognitive strategies in this model.

In addition to the two extensively used frameworks, other models and ideas that warrant consideration are Anderson's (2002) model of metacognition in learning and teaching, Efklides' proposal of metacognitive skills (2002) and Papaleontiou-Louca's (2003, 2008) overview framework of metacognition (Zare, 2012). In Anderson's (2002) model, metacognition demonstrates its role as strategy use, and it is comprised of five components: Preparing and planning for learning, selecting and using strategies, monitoring strategy use, orchestrating strategies and evaluation on strategy use. Based on Flavell's (1979) and Brown's (1987) framework, Efklides' (2002) metacognitive skills refers to the "conscious control process such as planning, monitoring of the progress of processing, effort allocation, strategy use and regulation of cognition" (Papaleontion-Louca, 2008, p. 15). Drawing upon an overview of metacognition, Papaleontiou-Louca's framework of metacognition includes metacognitive knowledge /metacognitive awareness (what one knows about himself or herself and others as a cognitive processor), metacognitive regulation (regulation of cognition), metacognitive

skills (conscious control such as planning, and monitoring) and metacognitive experience (experiences related to on-going cognitive activities). In essence, these more-recent frameworks or ideas on metacognition are variations of Flavell's (1979) and Brown's (1987) frameworks, and their illustration of metacognition via planning, monitoring and evaluation represents metacognitive regulation or metacognitive strategies (Zare, 2012). Such variation explains why scholars (e.g., Qin, 2018; Zhang & Qin, 2018; Zhang & Zhang, 2018; see also Livingston, 2003, Mahdavi, 2014; McCormick, 2003; Papaleontion-Louca, 2003, 2008; Purpura, 2003) have agreed that metacognition is comprised of metacognitive knowledge and metacognitive regulation or metacognitive strategies, and metacognitive strategies are composed of planning, monitoring and evaluating.

3.2.2 Metacognitive strategies in language learning strategies

In language learning and teaching (Anderson, 2012; Vandergrift, 2012), researchers in this area have reported that metacognitive strategies are the most important language learning strategies in a learner's successful learning (e.g., Griffiths, 2020; Oxford, 2017; Qin, 2018; Zhang & Qin, 2018; Zhang & Xiao, 2006; Zhang et al., 2019). Such importance of metacognitive strategies as language learning strategies accounts for the necessity of a review of this construct from the perspective of language learning strategies, a common practice for investigating test-takers' strategic competence in L2 assessment

3.2.2.1 Definitions and taxonomies of language learning strategies

The history of language learning strategies dates back to 1975 when Joan Rubin published her article on good language learners (Cohen, 2014; Griffiths, 2020; Oxford, 2017; Zhang et al., 2019). Since then, the research terrain has witnessed the ever burgeoning efforts with disagreements on some key issues including the definition and the taxonomies of the construct (Griffiths, 2020; Oxford, 2017; Plonsky, 2019; Zhang et al., 2019). Regardless of the disagreement, two well recognized taxonomy frameworks have been yielded: The Strategy System Model of Learning Strategies proposed by Oxford (1990) and O'Mally and Chamot' s (1990) Strategy Taxonomy Model (Dornyei, 2005).

Oxford's (1990) Strategy System Model of Learning Strategies

As "the most inclusive taxonomy of language learning strategies" (Zare, 2012, p. 165), Oxford's (1990) Strategy System Model of Learning Strategies presents language learning strategies as "specific actions taken by the learners to make learning easier, faster, more enjoyable, more self-directed, more effective, and more transferable to new situation" (p. 8). In the model, metacognitive strategies are regarded as actions beyond learners' cognitive devices, and they control learners' cognition by coordinating their learning process via centering, arranging, planning, monitoring and evaluating. Centering one's learning involves overviewing and linking with prior knowledge, paying attention, and delaying speech production to focus on listening; arranging and planning one's learning indicates organizing, setting goals and objectives, identifying the purpose of a language task, planning for a language task, and seeking practice opportunities; monitoring relates to self-monitoring and self-evaluating to check one's performance (Qin, 2018; Zhang, 2017; Oxford, 2011, 2017; D. Zhang & Zhang, 2019; L. Zhang & Zhang, 2018; Zhang et al., 2019).

O'Mally and Chamot's (1990) Strategy Taxonomy Model (1990)

Drawing from Brown's (1983) thoughts on metacognitive strategies, O'Mally and Chamot (1990) put forward the Strategy Taxonomy Model of Learning Strategies, which is the most influentially adopted model for exploring and categorising learning strategies (D. Zhang & Goh, 2006). According to O'Mally and Chamot (1990), learning strategies are "the special thoughts or behaviours that individuals use to help them comprehend, learn, or retain new information" (p. 1). In the model, metacognitive strategies are high order executive skills, encompassing planning, monitoring and evaluating. Planning involves previewing tasks for using relevant strategies and planning for the language used. Monitoring contains self-monitoring associated with self-checking the accuracy of one's task performance, and it is used for examining, verifying or correcting an individual's understanding of or performance in a specific task. Evaluating is post-task examination of individuals' performance with reference to a criterion associated with checking the outcome of the performance such as the completeness of the task and the accuracy of the language used. This criterion also has to do with the examination of individuals' strategy use.

3.2.2.2 Metacognitive strategies as a component of language learning strategies

In the research domain of language learning strategies, the two most influential models reviewed above entail various models, all of which accommodate metacognitive strategies as one of the essential language learning strategies (Dörnyei 2005, Griffiths, 2020; Thomas & Rose, 2019). For example, after comparing Oxford's model (1990) with that of O'Malley and Chamot's (1990), Dörnyei (2005) proposed his model of language learning strategies in which metacognitive strategies contain high-order strategies for analyzing, monitoring, evaluating, planning and organizing learners' learning process. In discussing language learner strategies, Cohen (2014) provided his understanding of metacognitive strategies as pre-assessment, preplanning, online planning, monitoring, and post-evaluation of language learning and use. In Oxford's (2017) more recent revised model of language learning strategies, metacognitive strategies are composed of paying attention to cognition, planning for cognition, obtaining and using resources for cognition, organizing for cognition, implementing plans for cognition, orchestrating cognitive strategy use, monitoring and evaluating cognition.

From the exposition above, it is evident that the seemingly different models "reflects relatively the same categorisations of language learning strategies without any fundamental changes" (Zare, 2012, p. 164) and the key elements of metacognitive strategies across these models are consistent: Planning, monitoring and evaluating (Qin, 2018; Oxford, 2011, 2017; D. Zhang & Zhang, 2019; L. Zhang & Zhang, 2018; Zhang et al., 2019).

3.2.3. Metacognitive strategies: Integration of metacognition with learning strategies

The above review indicates that the definitions and components of metacognitive strategies in the arena of language learning strategies overlap considerably with those in the research realm of educational psychology. Moreover, the essential elements of metacognitive strategies are widely recognized as planning, monitoring and evaluating (e.g., Griffiths, 2020; Qin, 2018; Oxford, 2017; D. Zhang & Zhang, 2019; L. Zhang & Zhang, 2018), which is consistent with the tripartite model of metacognition proposed by Wenden (1987), the first scholar who successfully integrated the research efforts in

metacogniton with those in L2 learning and teaching (Qin, 2018; Zhang, 2010; Zhang & Qin, 2018). In Wenden's (1987) model of metacognition, metacognitive strategies include planing individual's learning activities in accordance with their learning objectives prior to L2 learning; on-line monitoring in the individuals'learning process and post-learning evaluating of their learning process.

3.2.4. Working mode

The review of metacognitive strategies also shows the working mode of the construct displayed in the following characteristics.

3.2.4.1 Working independently and interactively in a simultaneous form

Metacognitive strategies do not work in isolation. Instead, they act both independently and interactively to exert impact on individuals' cognitive activities (Flavell, 1979, 1981). Although metacognitive strategies occasionally function independently, when learners perform a task, they often use more than one strategy at a time (Cohen, 2014), as metacognition is not a linear process which moves from preparing and planning to evaluating (Anderson, 2012). Rather, the strategies are likely to co-occur and the earlier strategies (e.g., planning) prompt later strategies (e.g., monitoring) recursively, affecting the preceding strategies and accordingly modifying the original plans set by the task-takers through a feedback loop (Schmitz &Wiese, 2006). A successful learner is believed to use clustering strategies rather than use the strategies in sequence (Takeuchi, 2019). Similarly, a test-taker's metacognitive strategies interact with one another, and with the factors of test tasks in the actual process of L2 assessment (Bachman & Palmer, 2010). The independent and the simultaneous clustering working mode of metacognitive strategies has been well summarised by Azevedo (2009) who commented that there is no strong assumption in the literature that the components of the metacognitive strategies are "hierarchically or linearly structured such that earlier phase must occur before later phases" (p. 87).

3.2.4.2 Task-dependent

The variation of metacognitive strategy use across different tasks has been agreed upon by researchers (e.g., Gu, 2019; Oxford, 2017; Cohen, 2018; Wang et al., 2015). For instance, Flavell (1979, 1981, 2000) contended that individuals' employment of their metacognitive strategies responds accordingly to the tasks given. Likewise, Brown

(1987) advocated that metacognitive strategies, the central executor or monitoring system of an information processing system, are task-dependent, implying that one's use of metacognitive strategies is contingent upon tasks. In echoing the two researchers' view, Kluwe (1987) postulated that in processing information, individuals' activation of their metacognitive strategies involves decision-making which will help them identify the task. Based on the identification, they will decide how to allocate their resources for that task, and determine the order of the steps for finishing the task. In the meanwhile, metacognitive strategies help the individuals to set the intensity or the speed at which they will work for that given task.

3.2.4.3 Mediating test tasks and test-takers' performance

Because of the above general characteristics, in L2 assessment, Bachman and Palmer (2010) underscored the role of strategic competence or metacognitive strategy use as a mediator between test-takers and test tasks, influencing test performance (Skehan, 2018). Similarily, Chapelle (1999) and Weir (2005) emphasised the seminal role of the metacognitive strategies in testing, arguing that metacognitive strategies must be treated as a mediator in the interactions between test-takers and test tasks. In fact, in L2 speaking assessment research, strategic competence has long been acknowledged to be an integral component of test-takers' speaking ability, mediating the relationships between speaking test tasks and speaking performance (e.g., Barkaoui et al., 2013; Bygate, 1987; Fulcher, 2014, 2015a, 2015b; Seong, 2014; Barkaoui et al., 2009). Concrete examples are Swain et al.'s (2009) study and that of Barkaoui et al. (2013). Both research teams inferred from their studies on EFL learners' strategic behaviours in the TOEFL speaking tests that test-takers' strategic competence, though not considerably affecting their test performance, was a mediator between test-takers and the test tasks.

It should be stressed here that in a specific study, a mediator refers to a variable that has correlations with both dependent variables and independent variables, explaining why and how the relationship between the two types of variables occurs. In correspondence to a mediator is a moderator as discussed in Chapter Seven and Chapter Eight. A moderator is a variable that affects the strength and the direction of the relationship between the two types of variables, though it does not necessarily have a correlation with both of them (Hayes, 2018; Karazsia, & Berlin, 2018; MacKinnon 2011). The

difference between a mediator and a moderator is shown in Figure 3.3.

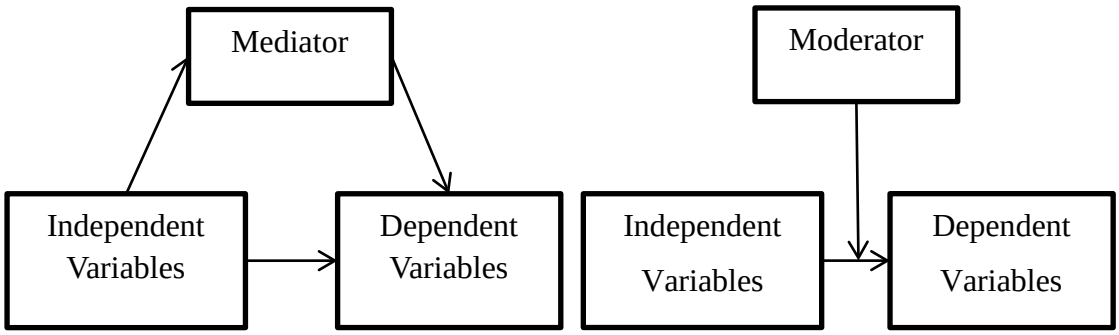

Figure 3.3 *Difference between a Mediator and a Moderator*

3.2.5. Definition, taxonomy and working mode of metacognitive strategies for this study

As reviewed above, typically, metacognitive strategies refer to planning, monitoring and evaluating. However, Chamot (2005) has argued that problem-solving should be considered as one component of metacognitive strategies based on her and her colleagues' (1999) Metacognitive Model of Strategic Learning (hereinafter referred to as the Chamot's Model). According to Chamot et al. (1999), the inclusion of problem-solving is due to its "usefulness and applicability to a broad range of learning tasks" (p. 11). Moreover, Chamot (2005) pointed out that almost all the metacognitive models that highlight metacognition in examining learning strategies include problem-solving as the fundamental component with planning, monitoring and evaluating: The Chamot's (1999) Model as noted above, Rubin's self-management model (2001) composed of planning, monitoring, evaluating, problem-solving and implementing; Anderson's (2002) Model of Metacognition (see Section 3.2.1.1), and the metacognitive model proposed by the National Capital Language Resource Centre (2003) in the USA in which problem-solving lies at the centre with the surrounding metacognitive strategies of planning, monitoring and managing and evaluating.

Since an exploratary approach was adopted in this study for the examination of the strategic competence, this four-component model was employed. Yet, given that metacognitive strategies were investigated in the context of L2 speaking assessment,

different from the normal non-testing learning setting where the Chamot's Model was devised, only the subcomponents consistent with test contexts in the model were under investigation. In light of this, the working definition and the taxonomies of the metacogntive strategies in this study are clarified in Table 3.1.

Table 3.1 *Definitions and Taxonomies of Metacognitive Strategies in This Study*

MS	**Taxonomies**	**Definitions**
Planning	Setting goals	Identify the purpose of the task
	Directed attention	Decide in advance to focus on particular tasks and ignore distractions
	Activate background information	Think about and use what you already know to help you do the task
	Prediction	Anticipate information to prepare and give direction for the task
	Organizational planning	Plan the task and content sequence
	Self-management	Arrange for conditions that help you learn
Problem-solving	Inference	Make guesses based on previous knowledge
	Substitute	Use a synonym or descriptive phrase for unknown words
Monitoring	Selective attention	Focus on key words, phrases, and ideas
	Deduction/induction	Consciously apply learned or self-developed rules
	Personalize/personal experience	Relate information to personal experiences
	Take notes	Write down important words and concepts
	Ask if it makes sense	Check understanding and production to keep track of progress and identify problems
	Self-talk	Talk to yourself to reduce anxiety by reminding self of progress, resources available, goals
Evaluating	Verify predictions and guesses	Check whether your predictions or guesses are correct
	Check goals	Decide whether goal was met
	Evaluating performance	Judge how well you do in the task

Note. MS = metacognitive strategies, this table is adapted from *The learning strategies handbook*, by A. U. Chamot, P. B., S. Barnhardt, P.B. El-Dinary and J.Robbins, 1999, pp. 15-18. Copyright 1999, Longman, White Plains, New York, USA.

In addition to the exploratory approach, the features of the Chamot's Model and its correspondence to Bachman and Palmer's (2010) strategic competence model also serve as rationales for the adoption of the four-component model as discussed below.

3.2.5.1 Features of the Chamot's Model

The Chamot's Model is built upon the combination of metacognition and learning strategies which suggests the dual identities of metacognitive strategies as a component of metacognition and as a constituent of the learning strategies. Furthermore, the model is based on extensive studies on language learning strategies, in which data were collected from effective L2 learners with various backgrounds ranging from elementary school students to those at the tertiary education level. The diverse backgrounds of learners and the variety of the learning tasks they performed evidence that the Chamot's Model is empirically grounded (Chamot, 2009). Additional supporting evidence has to do with the recursive nature of the model, its characteristics of being task-dependent, and its emphasis on the interaction between tasks and task-takers: The four metacognitive strategies are "not strictly sequential but may be used as necessary depending on the demands of task and the interaction between the task and the learner" (Chamot et al., 1999, p. 12). These characteristics not only well demonstrate the general features of metacognitive strategies, but also make feasible the establishment of the correspondence between the Chamot's model and Bachman and Palmer's (2010) strategic competence model, as both models involve the interaction between strategy use and tasks.

3.2.5.2 Correspondence between the two models

In L2 assessment, the interpretations of Bachman and Palmer's (2010) strategic competence model (see Table 2.1) show that the three components in the model (goal setting, appraising and planning) correspond to planning, monitoring and evaluating in the Chamot's Model (see Table 3.1). In terms of problem solving, Bachman and Palmer (2010) stated that their strategic competence model is derived from Sternberg's (1985, 1988) interpretation of meta-components of intelligence, which refer to planning, monitoring and evaluating individuals' problem solving. The statement indicates the necessity of connecting the three meta-components to problem solving. Also, as Seong (2014) commented, earlier definitions of strategic competence including Bachman and

Palmer's (2010) model are considerably influenced by Canale and Swain's (1980) model of communicative competence in which strategic competence is understood as problem-solving mechanisms. The indispensable role of problem solving in Bachman and Palmer's (2010) strategic competence model further rationalises the correspondence between the two models.

3.2.6 Measurement

Individuals' metacognitive strategy use is commonly measured via internal self-reports such as questionnaires and oral interviews (Craig et al., 2020; Nett et al., 2012; Ong, 2014; Veenman & van Cleef, 2019). Questionnaires are also termed as "inventories", "scales" and "survey" (Dörnyei, 2010), and the instrument has great popularity in empirical studies on metacognitive skills due to the following properties (Craig et al., 2020; Ong, 2014; Veenman & van Cleef, 2019):

a) It is the most cost-effective and efficient method to evaluate individuals' metacognitive activities;
b) It is easily administered on a large sample size;
c) It is used as the least intrusive method;
d) Data collected via questionnaires are easy to analyse, and hence are applicable in many statistical analyses;
e) The validity and the reliability of the instrument in measuring metacognitive strategies have long been extensively tested and verified.

In research on metacognitive strategies, questionnaires are often used with oral interviews, because this measure avoids the bias (Semerari et al., 2012; Schellings et al., 2011). Additionally, interviews are one of the major instruments for scrutinising participants' knowledge and application of metacognitive strategies. Among the three types of interviews (viz. open or unstructured interviews, semi-structured interviews and structured interviews), a semi-structured interview generates uncontrolled, open, improvised and true responses from participants, as it allows the respondents to answer questions freely without strict boundaries. Consequently, it permits researchers to explore all the aspects of research questions (Dörnyei, 2010). Henceforth, The integration of questionnaires with semi-structured interviews enables researchers to conduct complementary and in-depth data analysis (Sun, 2016), which rationalises the

employment of the two instruments to measure the Chinese EFL learners' use of metacognitive strategies: The Metacognitive Strategy Inventory (MSI) and the Semi-structured Interview Guide (SSIG).

In L2 assessment, thanks to its validity, Purpura's (1997) Metacognitive Strategy Questionnaire, though primarily focusing on reciprocal communications in non-testing context, has been adopted and adapted in quite a few studies for eliciting metacognitive strategies (e.g., Phakiti 2008; Zhang, 2017). The questionnaire is composed of four parts (assessing the situation, monitoring, self-evaluating and self-testing), and a six-point Likert scale ranging from 0 (never) to 5 (always) is used. Based on this questionnaire, Phakiti (2003) devised his metacognitive questionnaire on EFL reading test. Compared with Purpura (1997), Phakiti used fewer items and a five-point Likert scale, which makes the questionnaire more user-friendly. Besides, it is an off-line self-report suitable for reading testing conditions.

Since metacognitive strategies are considered as one form of language learning strategies, Oxford's (1990) Strategy Inventory of Language Learning (SILL) has also been extensively adopted in relevant empirical studies (Liu, 2018; Nett et al., 2012; Amani, 2014; Sun, 2016; Teng, 2016; Zhang & Xiao, 2006). The SILL aims at general learning strategy use, and therefore it comprehensively includes six types of strategies: Memory strategies, cognitive strategies, compensation strategies, metacognitive strategies, affective strategies and social strategies. A five-point Likert scale ranging from 1 (never use it) to 5 (often use it) is employed in the inventory. Due to its generalness, the SILL is unlikely to be applied in a specific context directly without any modifications (Amani, 2014; Schellings et al., 2013).

With respect to L2 speaking, questionnaires that examine metacognitive strategy use in this context are extremely scant. To my best knowledge, only one such a questionnaire is available: Metacognitive Awareness Inventory in Listening and Speaking Strategies (MAILSS) developed by Zhang and Goh (2006). In the MAILSS, strategies for speaking and listening are divided into four groups: Use-focused learning strategies, form-focused learning strategies, comprehension strategies, and communication strategies. The frequency of metacognitive speaking strategy use is rated on a scale from 'Never'

(1) to ‘Very Often’ (5). Although the MAILSS can be used to investigate individuals’ metacognitive speaking strategies, it is not developed especially for speaking, and it focuses on EFL learners’ development of their metacognitive awareness in the non-testing conditions. Because of the limitations, the inventory has not been widely applied (Craig et al., 2020; Ghapanchi & Taheryan, 2012).

In terms of the semi-structured interview guide, Zhang (1999) originated the Scheme for Eliciting Subjects’ Metacognitive Knowledge through Semi-structured Interviews to investigate the metacognitive strategies used by Chinese EFL readers. The scheme is in Chinese and it includes questions regarding how Chinese EFL learners perceive their use of metacognitive strategies and how they solve problems in reading.

The properties of the above questionnaires and the interview guide in validity, the participants on whom the instruments are used, the language skills investigated via the instruments, and the contexts (testing or non-testing) where they are applied account for why these instruments serve as the original sources of the development of the MSI and the SSIG (see Section 4.5.1.1 for detailed information of the instruments).

3.3 Task Complexity

As noted earlier, the test task characteristic of input format in the four integrated speaking test tasks (see Table 2.3) indicates the varying degrees of task complexity. However, in the research on speaking tests, the investigation into test tasks within Bachman and Palmer’s (1996, 2010) model of language use tasks is extremely challenging (Iwashita et al., 2001; Fulcher & Reiter 2003). Because of this, I investigated the four tasks in Robinson’s (2015) Triadic Componential Framework, a well acknowledged approach to task research (Abdi Tabari, 2018; Choong, 2014; Lee, 2018), through establishing a correspondence between the framework and the input format in Bachman and Palmer’s model. The subsequent subsections vindicate the establishment.

3.3.1 Approaches to characterising tasks

According to Iwashita et al. (2001), Bachman and Palmer’s (1996, 2010) Language Use Tasks Characteristics Model or test task characteristics mode (see Figure 2.2) reflects

one of the three approaches to characterising test tasks in concert with the interactional approach and the information-processing approach. The interactional approach stems from the interactionists' work (see Section 2.2.2.2) and is used to determine the interactional characteristics of tasks and their impact on task performance in a dyadic conversation. The information-processing approach is best known in the work of Skehan (2016, 2018) regarding task difficulty and in the work of Robinson on task complexity (2011a, 2011b, 2015). In this approach, the impact of the cognitive characteristics of tasks on performance and the link between task difficulty/complexity and task performance can be examined. Despite the difference in the three approaches, there is a broad agreement that task characteristics of a specific task indicate task complexity, influencing task performance, and among the characteristics, input is the basic source of task complexity (e.g., Ellis et al., 2019; Liu & Li, 2012; Skehan, 2018; Wigglesworth & Frost, 2017). Bachman's (2002) statement that tasks differ in difficulty also suggests that the complexity of a specific task is ingrained in its task characteristics. The close link between task characteristics and task complexity accounts for, to some degree, why the three approaches are equally adopted in the research on task complexity in L2 speaking assessment as reviewed hereunder.

3.3.2 Approaches to examining task complexity in L2 speaking assessment

In L2 speaking assessment, task complexity is regarded as a construct that is too complex to be defined and measured (Bachman, 2002; Elder et al., 2002; Fulcher & Reiter, 2003; Iwashita et al., 2001; Wigglesworth & Frost, 2017). Hence, various approaches have been adopted to examine the complexity of speaking test tasks. Fulcher and Reiter (2003) categorised these approaches as three approaches to task conceptualisation, which were summarised by Iwashita et al. (2001) as approaches to characterising tasks as reviewed above. In the three approaches, the information-processing approach based on Skehan's (1996, 1998) conceptualisation of task difficulty is used the most (Elder et al., 2002; Iwashita et al., 2001; Weir et al., 2006). The reason is that the interaction approach is more related to classroom learning than language tests, and Bachman and Palmer's (1996, 2010) model of test characteristics presents an "unordered check-list", making it difficult for researchers to refer to in examining test tasks (Iwashita et al., 2001; Fulcher & Reiter 2003). The drawback of Bachman and Palmer's (1996, 2010) model on test tasks makes its wide applicability hard to achieve,

and hence, the information processing approach has overshadowed the other two, playing a dominating role even in today's research terrain of L2 assessment (e.g., Abdi Tabari, 2018; Choong, 2014; Gan, 2012; Lee, 2018).

Consequently, irrespective of the profound impact of Bachman and Palmer's (1996, 2010) Language Ability Model on language assessment, language testers and researchers do not commonly use the model. Rather, the information processing approach represented by Skehan's Limited Attention Capacity Model (2016, 2018) and Robinson's Cognition Hypothesis (2015) is typically adopted in the exploration of test tasks. The situation necessitates my efforts to investigate the integrated speaking tests in an approach that not only corroborates with Bachman and Palmer's model within the theoretical framework but also makes the investigation feasibly operationalizable and simultaneously well-grounded (Pallotti, 2019).

In the information processing approach, especially Skehan's (1996; 1998) task difficulty approach, task difficulty is conceptualised in accordance with three sets of task characteristics: Code complexity, cognitive complexity and communicative stress which are associated with language, the thinking and the performance conditions required on test-takers to finish the tasks respectively (Bachman, 2002; Iwashita et al., 2001; Fulcher & Reiter 2003). This task difficulty approach is criticised by Bachman (2002) as "essentially an artifice of performance and not a characteristic of assessment task themselves" (p. 465). He argued that examining test task characteristics in this approach confounds test tasks with test-takers, as the difficulty features identified in the approach are essentially a combination of ability requirements on test-takers and test task characteristics. In other words, in Skehan's approach, a clear-cut distinction between test-takers and test tasks is hard to make (Robinson & Gilabert, 2007). Moreover, Bachman (2002) and Fulcher and Reiter (2003) attributed the unsuccessfulness and the unexpected mixed findings of the task difficulty studies in L2 assessment to the employment of this approach. Essentially, Bachman's (2002) criticism on the task difficulty approach suggests the distinction between task complexity and task difficulty as reviewed in the following subsection.

3.3.3 Task complexity and task difficulty

Although Skehan's (1998) task difficulty approach was adopted extensively in early studies on L2 assessment, ever since the term "task complexity" was consistently used by Robinson (2001, 2005, 2007, 2011a, 201lb, 2015) to differentiate task difficulty and task complexity, the task complexity approach has continued to gain ground in the research on task characteristics (Pallotti, 2019), especially in the studies on L2 speaking tasks (e.g. Adams & Alwi, 2014; Levkin & Gilabert, 2012; Tajeddin & Bahador, 2012). Because of such influence, task complexity approach and task difficulty approach play an equally dominant role in the task research area (Ellis et al., 2019; O'Grady, 2018; Lee, 2018; Rahimi, 2016), and "have influenced a wide range of studies" (Skehan, 2016, p. 35).

Unlike Skehan (1996, 1998, 2012, 2016, 2018) who used code complexity, cognitive complexity and communicative stress to interpret task characteristics (Bachman, 2002), Robinson (2011a, 2011b, 2015) and Robinson and Gilabert (2020) argued that tasks should be considered from the perspectives of task difficulty, task complexity, and task conditions, which is well illustrated by Robinson's (2015) Triadic Componential Framework (see Figure 3.4). In the framework, task difficulty is conceptualised as learner factors closely connected to learners' perceptions of task demands, contributing to the variance in task performance between task-takers. In contrast, task complexity links to task factors, indicating the intrinsic cognitive demands posed on task-takers. Task complexity is a relatively stable and inherent task characteristic, and it is a series of objective task characteristics or variables accounting for the variance in task performance within task-takers. Task condition concerns the conditions under which task-takers perform tasks including participation (e.g., whether a task is a monologue or a dialogue) and participants (e.g., whether participants engaging in the tasks are familiar or unfamiliar with the topic) (e.g., Lee, 2018; Rahimi, 2016; Rahimi & Zhang, 2018).

1. *Task Complexity* (Cognitive factors)	2. *Task Condition* (Interactive factors)	3. *Task Difficulty* (Learner factors)
(a) *Resource-directing variables* making cognitive/conceptual demands	(a) *Participant variables* making interactional demands	(a) *Ability variables* and task-relevant resource differentials
+/- here and now +/- few elements -/+ spatial reasoning -/+ causal reasoning -/+ intentional reasoning -/+ perspective-taking	+/- open solution +/- one-way flow +/- convergent solution +/- few participants +/- few contributions needed +/- negotiation not needed	h/l working memory h/l reasoning h/l task-switching h/l aptitude h/l field independence h/l mind/intention-reading
(b) *Resource-dispersing variables* making performative/procedural demands	(b) *Participant variables* making interactant demands	(b) *Affective variables* and task-relevant state-trait differentials
+/- planning time +/- single task +/- task structure +/- few steps +/- independency of steps +/- prior knowledge	+/- same proficiency +/- same gender +/- familiar +/- shared content knowledge +/- equal status and role +/- shared cultural knowledge	h/l openness to experience h/l control of emotion h/l task motivation h/l processing anxiety h/l willingness to communicate h/l self-efficacy

Figure 3.4 *Robinson's (2015) Triadic Componential Framework*

Source: Task complexity, the Cognition Hypothesis and secondlanguage learning and performance, by P. Robinson, and R. Gilabert, 2007. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(3), p. 164. Copyright 2007 by De Gruyter Mouton.

3.3.4 Correspondence between the two frameworks

The distinction between task complexity and task difficulty in Robinson's (2015) Triadic Componential Framework essentially serves as a response to Bachman's (2002) call for studying test tasks without mixing them with test-takers when commenting on Skehan's (1998) task difficulty approach. In addition to the clear-cut differentiation, Robinson also regards a task as a multi componential construct, believing that tasks are the combination of a set of task characteristics. Such a view is consistent with Bachman and Palmer's (2010) definition of test tasks. The consistency between Robinson's (2015) Triadic Componential Framework and Bachman and Palmer's (2010) Language Use

Task Characteristics Model make it tenable to establish a correspondence between the two frameworks. Another piece of evidence that supports the correspondence comes from the information-processing elements involved in both frameworks.

As reviewed above, as early as in 2001, Iwashita and her colleagues contended that both Skehan's task difficulty approach and Robinson's task complexity approach are fundamentally categorised as information-processing approaches. This contention has been acknowledged by many scholars (e.g., Ellis, 2015; Ellis et al., 2019; Sasayama, 2015, 2016; Wang & Zhang, 2019), and even Skehan (2015, 2018) and Robinson (2011a, 2011b, 2015) explicitly associated their approaches with the information processing process. According to Skehan (2016, 2018), such an information-processing approach makes it reasonable to establish the connection between cognitive demand of tasks in speech production to task-takers' strategic competence in Bachman and Palmer's (1996, 2010) models of language ability. The connection is further evidenced in Kormos' (2006, 2011) L2 speech production model (see Section 3.4.2). In line with the interaction between task demand and strategic competence illustrated in the theoretical framework (see Figure 2.3), it is obvious that the two constructs (tasks and strategic competence) connected by this information-processing approach suggests that Robinson's Triadic Componential Framework (2015) corresponds to Bachman and Palmer's (2010) model on test task characteristics.

To summarise, the above literature review shows three fundamental properties of Robinson's (2015) Triadic Componential Framework:

a) There is a clear distinction between task complexity and task difficulty with the former relating to tasks and the latter pertaining to task-takers;
b) A task encompasses multiple features;
c) Task factors reflected in the framework have close associations with task-takers' strategic competence in their performance on L2 speaking tasks.

Taken together, it is justifiable to establish the correspondence between the component of task complexity in Robinson's (2015) framework to Bachman and Palmer's (2010) Language Use Task Characteristics Model. Accordingly, the test task characteristics of input format in the four (see Table 2.4) integrated speaking test tasks can be investigated

within Robinson's framework. As stated previously, such correspondence enables researchers to mitigate the inapplicability of Bachman and Palmer's model with a one that is extensively recognised in the task characteristics studies, and concomitantly to maintain the essence of tasks: A task is a synthesis of diverse properties independent of task-takers.

3.3.5 Task complexity in Robinson's (2015) Triadic Componential Framework

As the most detailed and operational framework that distinguishes the componential dimensions of task characteristics and predict their effects on performance to date (Lee, 2018, 2019), Robinson's (2015) Triadic Componential Framework is based on task designing and sequencing in accordance with task complexity for the duplication of real-world tasks (Lee, 2018). It represents his (2001a, 2001b, 2003, 2005, 2007, 2011a, 2011b, 2015) Cognition Hypothesis, the tenets of which hold the view that task complexity refers to the cognitive demands of tasks, which is manipulable and task-dependent, and it is "the result of the attention, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner" (Robinson, 2011a, p. 29). In line with this hypothesis, Robinson (2015) proposed that in the Triadic Component Framework (see Figure 3.4), a series of task factors or variables can be manipulated along the resource-directing and resource-dispersing dimensions for various levels of task complexity. His proposal of resource-directing variables distinguishes task complexity in conceptual or linguistic demands on task-takers. Those variables include: (a) Whether tasks require learners' refer to events now (Here-and-now) versus events in the past (There-and-Then); (b) whether learners have to refer to few elements or many elements (+/- few elements); (c) whether task-takers have to give reasoning about other people's intentions, beliefs and desires and the relationship between them (+/- intentional reasoning); (d) the availability of supportive information related to spatial location (+/- spatial reasoning); (e) task requirement on how to process information (simple information transmission versus reasoning the intentions of other people); and (f) the perspective that task-takers take to accomplish tasks (e.g., the first person perspective or the third person perspective). Contrary to resource-directing variables, resource-dispersing variables impose informative or procedural demand on task-takers' attentional and memory resources, and these variables indicate: (a) Whether planning time is provided or not (+/- planning); (b) if

background knowledge or familiarity with the task content is available (+/- prior knowledge); (c) whether few steps or many steps are needed to complete a given task (+/- few steps); (d) whether task-takers have to take the steps in sequence or if the steps are independent from one another (+/- independency of tasks); (e) there is only one thing to be done or multiple things to be done before a task can be finished (+/- few steps); and (f) whether the structure of a task given is clear or not (+/- task structure). Robinson further proposed that increasing task complexity along the two dimensions affects task-takers' performance. Regarding the synergetic effects of the variation in task complexity via the manipulation of multiple task factors simultaneously along the two dimensions, Robinson stated that the effects of increasing task complexity along the resource-directing dimension may be weakened or negated by the increase of task complexity along the resource-dispersing dimension. The comprehensiveness of Robinson's framework makes it more feasible, compared with Skehan's model of task difficulty, to explore the effects of the manipulation of multiple task complexity variables simultaneously on task performance, and to predict the relationships between these variables (Abdi Tabari, 2018).

Although Robinson's framework has a prominent influence on task complexity research, some scholars (e.g., Kuiken & Vedder, 2007; Skehan, 2016, 2018) have questioned the validity of the "comprehensive criteria". They argued that the framework is hard to be researched and operationalised in actual empirical studies (Salimi & Dadashpour, 2012), a complaint similar to the one that opposes Bachman and Palmer's framework of test task characteristics as discussed earlier. Moreover, scholars such as Skehan (2016, 2018) and Ellis (2017) question the validity of the manipulation of task complexity along the two dimensions: The manipulations of task complexity may not elicit the assumed varying degrees of complexity. To solve the question, researchers including Robinson (2005), Révész (2011, 2014), Révész, Michel and Gilabert (2016), Sasayama (2015, 2016), Kim, Payant and Pearson (2015), and Lee (2018, 2019) have worked out a few well-recognised solutions (Skehan, 2016, 2018) to validate the manipulation of task complexity (see Section 3.3.6). The validation also rationalises why I examined task complexity of the four speaking tasks within Robinson's (2015) framework based on its correspondence to Bachman and Palmer's framework.

3.3.5.1 Task complexity variables in the integrated speaking test tasks

In accordance with the correspondence between Robinson's (2015) and Bachman and Palmer's (2010) frameworks, the task characteristic of input format in the four integrated speaking test tasks can be translated into four task complexity variables. From Table 2.1 and Figure 3.3, it can be seen that three variables along the dimension of resource-dispersing are identified in the four tasks: Planning time, steps involved and prior knowledge, which correspond to the input characteristics of time (preparations), input for interpretation/length and topic knowledge respectively. It is worth noting that irrespective of the difference between Skehan and Robinson in their proposals of task complexity, Robinson (2015) commented that the two scholars agree that manipulating variables along the dimension of resource-dispersing in Robinson's framework affects task performance. The agreement is obviously another argument for a need to examine the integrated speaking test tasks using Robinson's framework. In terms of the input characteristic of prompt, existing literature shows that a few scholars (e.g., Kim, Payant & Pearson, 2015) have classified it as reasoning demand within Robinson's task complexity model. Despite this, I considered it as task type in line with the majority of the researchers (e.g., Barkaoui, 2015; Barkaoui et al., 2013; Kaivanpanah, Yamouty & Karami, 2012; Pan & In'nami, 2015; Swain et al., 2009; Yi, 2012; Youn & Bi, 2019). In the following subsection, literature on task type for classifying the prompts of the four tasks is reviewed.

3.3.5.2 Task types of the integrated speaking test tasks

As one of the task characteristics, task type reflects task complexity and a source of variation in test performance (Gan, 2012; Kim, 2009; Révész et al., 2016; Tavakoli & Rasekh, 2011; Madarsara & Rahimy, 2015; Weir, 2005). According to Foster and Skehan (1996), various task types require different cognitive load on task-takers and affect task complexity. In speaking, if other key factors are equal, task type variation will exert a consistent impact on speakers no matter whoever the speakers are and whatever their speaking context will be (Samuda & Bygate, 2008). Foster and Skehan (1996) and Skehan and Foster (1997, 1999) categorised tasks into three types: Personal information exchange tasks, narrative tasks and decision-making tasks. As a personal task relates to personal information, it is considered as the easiest one, while a decision-making task is regarded as the most difficult since it is about the input of a lot of new information and

evaluating and defending one's opinion on a particular issue, which means the extra weight on one's cognitive load.

In L2 assessment, based on Bygate' (1987) classification of task types, Luoma (2004) categorises speaking tasks as structured tasks and open-ended tasks. Structured speaking tasks "are the speaking equivalent of multiple-choice tasks" (p. 50), and are beyond the scope of this research, so a further review of this type of speaking tasks is not presented. Open-ended speaking tasks, on the other hand, are used to check individuals' speaking skills, which are comprised of eight types of tasks: Description, narration instruction and comparison, explanation, justification, prediction, and decision. The delineation on task types shows that Skehan and Foster's (1996) opinion on the three task types echoes that of Luoma (2004) and Bygate (1987). Given such consistency and considering the input characteristic of prompt reflected in the four integrated speaking tasks, the task types of these speaking tasks are categorised as a narration task for Task 3, a justification task for Task 4 and Task 6 and a decision plus justification task for Task 5.

3.3.5.3 Task complexity of the integrated speaking test tasks

The above review of the task characteristics in the four integrated speaking test tasks within Robinson's (2015) Triadic Component Framework and from the perspective of task type indicates the variability in task complexity of the four tasks. Table 3.2 shows the variability.

Table 3.2 *Task complexity in the Four Speaking Test Task*

Task	**Prior knowledge**	**Steps Needed**	**Planning Time**	**Task Type**
Task 3	Campus-life Situation	R-L-S	30s	Narration
Task 4	Academic Lecture	R-L-S	30s	Justification
Task 5	Campus-life Situation	L-S	20s	Decision-making Justification
Task 6	Academic Lecture	L-S	20s	Justification

Note. R = reading, L = listening, S = speaking, s=seconds

3.3.6 Measurement

As articulated above, task complexity indicates the cognitive demand of a specific task on task-takers. The manipulation of task complexity along the two dimensions of Robinson's (2015) Triadic Componential Framework is assumed to change the cognitive load of the task. When task-takers perform a complex task in opposition to a simple one, it is assumed that the task will impose greater challenge on their working memory with more working load, and hence affect the task-takers' performance. Nonetheless, in reality, the variance in task performance on the two tasks (the complex one versus the simple one) cannot be simply attributed to the variation in task complexity, and may be caused by other factors such as individual difference (Lee, 2018, 2019; Robinson, 2001a). Due to this, it is agreed that task complexity should be established independently before task-takers begin their task performance (e.g., Lee, 2018, 2019; Révész, 2011, 2014, 2015; Révész et al., 2016; Sasayama, 2015, 2016). However, in many prior empirical studies, researchers typically manipulate task variables along certain dimensions that are assumed to create a complex task in contrast with a simple task without proving that these tasks induce more or less cognitive demand on task-takers. Consequently, the validity of task complexity is problematic (Révész, 2011, 2014; Révész et al., 2016; Sasayama, 2016). To address this issue, researchers (e.g., Révész, 2011, 2014; Sasayama, 2015, 2016) have worked out some validated methods to measure the cognitive load of tasks as noted earlier (Lee, 2018, 2019, Skehan, 2016, 2018). According to Révész (2011, 2015), and Sasayama (2016), several methods have been identified as valid to measure independently the cognitive load of tasks, which include: (a) Self-rating scales/questionnaires, (b) subjective time-estimation, (c) dual-task methodology, (d) psycho-physiological techniques; and (e) expert judgment.

Self-rating scales/questionnaires were originally employed by Robinson (2010a) to investigate whether a complex task can elicit a high rating of perceived task difficulty by task-takers. The research method is motivated by the assumption that task-takers can assign a numerical value to the mental efforts they make for performing the tasks given, and they can also give a similar value to the difficulty of the tasks based on their perceptions. The positive correlation between the two values indicates task complexity assumed via the manipulation of task factors along a certain dimension does generate the assumed cognitive load on task-takers. Subjective time-estimation requires task-

takers to estimate how long it takes them to perform the given tasks (Baralt, 2010, 2014; Sasayama, 2015, 2016). It is assumed that task-takers spend a different amount of time on a given task in line with task complexity. In the dual-task methodology, task-takers are required to perform two tasks simultaneously: A primary task and the second task (a simple cognitive activity that demands sustained attention on task-takers). It is assumed that a task-taker's performance on the simpler task reflects the cognitive load of the primary task with reference to the task-taker's reaction time and accuracy in task performance. The psycho-physiological techniques include measuring task-takers' heart activities and recording their eye activities through eye tracking. Finally, the method of expert judgment is inviting experts to measure the cognitive load of a given task on a rating questionnaire according to their expertise on tasks (Lee, 2018, 2019; Révész, 2011, 2014; Révész et al., 2016; Robinson, 2001; Sasayama, 2016).

Sasayama (2015, 2016) argued that only using the above techniques, researchers could not explain why the measured-to-be complex task is more complex than the measured-to-be simple task, and they cannot explain if there are other factors (e.g., task design features) impacting the cognitive task complexity. To address these issues, she (2015) pointed out that multiple methods should be employed to validate the assumed task complexity. In light of this, a few researchers have tried introspective methods such as the stimulus recall by Kim et al. (2015) and the interviews by Tavakoli (2009).

Among the above methodologies, self-rating scales/ questionnaires, expert judgment and interviews were administered on the Chinese EFL learners and teachers to measure task complexity of the four integrated speaking tasks in this study (see Section 4.5.1 for detailed information on the instruments). Although the three techniques are subjective, the simultaneous employment of the three measures is extremely critical in that it is the task-takers, the Chinese EFL learners, who engaged in the cognitive process of test task performance, and it is their language ability that was under examination (Sasayama, 2015). Moreover, the combination of various measures of cognitive complexity rather than only one can make these measures complement each other; hence, the findings can be triangulated into a more accurate and thorough understanding of the cognitive task complexity in the four integrated speaking test tasks (Lee, 2018, 2019; Révész, 2011, 2014; Révész et al., 2016; Sasayama, 2015, 2016). Also, since task complexity measured

with the self-rating scales is only a rating of the synergetic effects of the simultaneous manipulation of task complexity factors involved in the four tasks, the use of the semi-structured interviews help to examine the independent effects of these factors on the Chinese EFL learners' speaking performance. Henceforth, data from the self-rating scale and from the semi-structured interviews can converge and align with each other to answer the research questions related to task complexity comprehensively (Creswell & Creswell 2018; Creswell & Guetterman; 2019; Harris & Brown, 2010).

On the other hand, the dual-task methodology, time-estimation and psycho-physiological techniques, though compared with the above three measures are more objective, were not employed in this study for the following reasons: (a) As the Chinese EFL learners had to perform the four tasks on computers within a limited time in a testing context, dual-task methodology and time-estimation might cause distraction to these test-takers; and (b) the psycho-physiological techniques such as recording one's eye tracking and heart activities are inaccessible due to the resource constrain.

3.4 Speaking and Speaking Performance

Thus far, within the theoretical framework of non-reciprocal language use, a multidisciplinary review of metacognitive strategies and test task characteristics has been conducted, and the definitions, taxonomies and the working mode of the Chinese EFL learners' strategic competence, and the cognitive complexity involved in the four integrated speaking test tasks have been clarified. Given that the theoretical framework is a general framework applicable in listening, reading, writing and speaking (Bachman & Palmer, 2010), further investigation on how strategic competence and task complexity work in speaking, and how the assumed interaction between the two variables within the theoretical framework affect performance in speech production is imperative. Therefore, the review of the two research variables in the specific context of speaking is presented in the subsequent subsections.

3.4.1 Models of speech production

Speaking is a very intricate productive skill (Newton & Nation, 2020; Yahya, 2019), and it is "a process of oral language production" (Tarone, 2005, p. 485). In the process, a speaker is an information processor (Hughes, 2017; Hughes & Read, 2017;

O'Sullivan, 2011, Bachman & Palmer, 2010) who receives and processes information for systematic utterances to express meaning that occurs in the real-time situations (Yahya, 2019). In the four language skills, speaking is "usually viewed as the most complex and difficult skill to master" (Tarone, 2005, p. 485). Speaking in a foreign language is more complicated and complex because speaking is done in real-time, and therefore speakers' abilities to plan and process involving metacognitive strategy use before their oral production are taxed greatly (Luoma, 2004).

In the current literature on speaking, models generated in the research field of psycholinguistics are widely recognized and applied (e.g., Kormos, 2006, 2011; Skehan, 2009, 2016; Sun, 2016; Xu, 2015; Yahya, 2019), among which Levelt's (1989) model of monolingual speech production has become "one of the most comprehensive and widely used theoretical frameworks" (Sun, 2016, p. 27). Consequently, it has been accredited as "the most influential model for monolingual speech production" (Xu, 2015, p. 38). Based on this model, De Bot's (1992) proposed his L2 speech production model, followed by many similar research efforts (e.g., Poulisse & Bongaerts, 1994; Towell et al., 1996). More recently, integrating Levelt's (1989) L1 model and existing L2 speech models, Kormos (2006, 2011) mapped out the Bilingual Speech Production Model. Compared with previous L2 speech models, this model is "more elaborate and more targeted" and has been acknowledged as the major bilingual model concerning speech production (Wang & Liu, 2018, p. 397). Hence, Kormos' bilingual model has been employed in many empirical studies on L2 speaking (e.g., Kormos, 2011; Xu, 2015; Yahya, 2019).

Considering its solid theoretical grounding and strong empirical support (Declerck & Kormos, 2012; Kormos, 2006, 2011; Wang & Liu, 2018, Xu, 2015; Yahya, 2019), I framed the present study in Kormos' (2011) Bilingual Speech Production Model for a deep look into the relationships between the research variables in the specific context of the integrated speaking test.

3.4.2 Kormos' (2011) Bilingual Speech Production Model

Due to the complexness of speaking as noted above and the characteristics of Kormos' (2011) Bilingual Speech Production Model, the review of the model is conducted at four

levels: General introduction, four stages involved, problem-solving mechanisms and attentional control.

3.4.2.1 General Introduction

As illustrated in Figure 3.5, Kormos' (2011) Bilingual Speech Production Model is modular, and it consists of separate encoding modules: A conceptualiser for planning message, a formulator for linguistically encoding the message, an articulator for articulating the encoded message as sounds. In addition, the model also encompasses a large knowledge store (a speaker's long-term memory) which provides the speaker with the information needed, a speech comprehension system which receives the speaker's actual discourse for inspection via monitoring, and an audition component (an acoustic-phonetic processor) that helps the monitor to check the produced utterance. The monitor is in the conceptualiser, monitoring the outputs of the conceptualiser, the formulator, and the whole process of speech production.

A speaker's long-term memory has a hierarchical level underpinned by activation spreading (Bygate, 2011; Wang & Liu, 2018; Xu, 2015; Yahya, 2019). Activation spreading is a model from the brain research area, and the working mode of the model is that neurons or interconnected cells interact with one another, establishing a neural network, through which simple signals of "activation" are exchanged within these neurons. The long-term memory accommodates the speaker's episodic memory (the speaker's memory on the episodic events that he or she experiences), and mental lexicon which is further broken down into three subcomponents: Conceptual knowledge, lemmas and lexemes. Conceptual knowledge refers to the knowledge that the speaker has acquired through his or her experience or the episodic memory; lemmas are syntactic words such as nouns and adjectives, and they are immediate from of words which contain the meaning and the syntax of a lexical entry or a word. On the other hand, the morphological and phonological information of a lexis entry is stored in lexemes. Moreover, a speaker's long-term memory also includes the speakers' syllabary, which stores a series of articulatory movements that the speaker can use for generating the syllables of a specific language. The speaker's declarative knowledge of L2 rules in syntax and phonology is also part of his or her long-term memory (Kormos, 2006, 2011; Xu, 2015; Yahya, 2019).

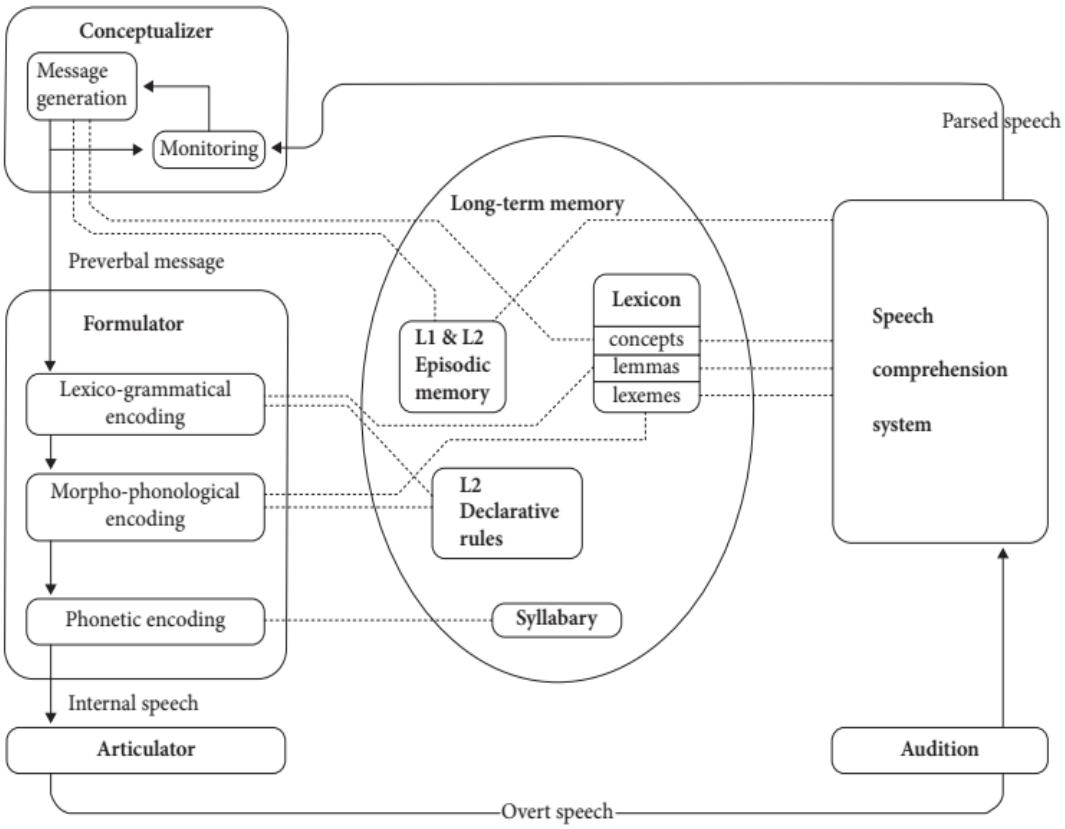

Figure 3.5 *Kormos' (2011) Model of Bilingual Speech Production*

Souce: Speech production and second language acquisition, by J. Kormos, 2006, p.168. Copyright 2006 by Routledge.

In correspondence to the modules are the four stages in L2 speech production: Conceptualisation in which the speaker plans what is going to speak; formulation where the speaker encodes linguistically the intended message; articulation through which the speaker executes his or her speech sounds by controlling the articulatory muscles, converting the phonetic plan generated by the formulator to overt speech; and monitoring with which the speaker checks and notices errors for possible modifications and corrections to make his or her utterance in light of the input or a specific speaking task. It is assumed that the four stages comprise the hierarchical level of the speaking process system, within which the information is also transmitted via activation spreading. Similar to L1 speaking process, L2 speech production also works incrementally, indicating that a fragment of a specific module's characteristic input can start the encoding procedures in that module. An example of this working mode is that when the first syllable of a word is encoded phonologically in the formulator, the

articulation of the word can begin in the articulator. As the entire speaking process requires a speaker's attentional control, each of the stages only works serially (Kormos, 2006, 2011; Xu, 2015; Yahya, 2019).

3.4.2.2 Four stages

Conceptualisation

As the starting stage of speech production, conceptualisation involves the activation of relevant concepts to be encoded, and the decision on the language used for expressing the concepts. Conceptualisation is realised via planning at the macro-planning level and the micro-planning level. In macro-planning, a speaker will select the information to be encoded and the order in which the information will be conveyed in accordance with his or her communication intension reflected by the input in the form of speech acts. Speech acts refer to the actions that the speaker takes such as directing, requesting, and apologising. With reference to task complexity, it is apparent that speech acts indicate the types of speaking tasks given to the speaker. In other words, during the macro-planning stage, L2 speakers will decide on the content of their speech in light of the speaking tasks given, for which the speakers' knowledge on relevant concepts stored in their mental lexicon will be activated.

After the speech content is specified, micro-planning begins, during which appropriate language is selected in accordance with the situation or the context where the intended message is conveyed. A concrete example of the language use appropriateness is that if a task is regarding something in the past, the speaker has to consider the correct tense used in the intended message. In addition, sociolinguistic factors also need to be taken into account. For instance, the language used for an academic purpose like a lecture differs from the language needed for daily talks between friends. The choice of language is indicated in the form of language cues which are added to the activated information on the concepts.

The output of conceptualisation is a preverbal plan. It specifies the concepts that reflect the speaker's intended messages, and contains the language cues which demonstrate what kind of language he or she selects to express the concepts in the intended discourses. Although the preverbal plan is not linguistic, it is linguistically available, as it contains

all the information needed to convert the speaker's speech intension to language (Kormos, 2006, 2011; Levelt, 1989; Skehan, 2018; Xu, 2015; Yahya, 2019).

Formulation

Formulation is essentially linguistic encoding of the preverbal plan at three levels: Lexicon encoding, syntactic encoding and phonological encoding. Lexicon encoding is to match the concept specification and language cues with an appropriate lexis entry in the speaker's long-term memory. The match commences with the activation sent by the conceptual specification to the lemmas in the speaker's mental lexicon and the lemmas whose properties corroborate with the preverbal plan will be selected. The syntactic features of these lemmas are then activated, which will trigger the ensuing procedure of syntactic building in which the speaker uses the syntactic encoding mechanism to arrange the phrases and clauses with the activated words and the syntactic features of these words. As the speaker's L2 knowledge is rarely complete, declarative knowledge on grammatical rules of the L2 in his or her long-term memory will be activated in the syntactic encoding, which is followed by the phonological encoding. The phonological forms of the selected words or lexemes are activated in the phonological encoding, which also involves the syllabification, the pitch and the length of intonational phrases composed of one word or a few words. After the linguistic encoding process at the three levels regarding morphology, syntax and phonology, a phonetic plan that indicates the internal speech is generated in the formulator (Kormos, 2006, 2011; Levelt, 1989; Skehan, 2018; Xu, 2015; Yahya, 2019).

Articulation

In articulation, the speaker's articulator retrieves and executes the phonetic plan via the articulatory organs, turning the internal speech into an actual discourse or the overt speech. The discourse is then received as the input by the audition component where the articulated utterance is recognised. Subsequently, the meaning of the utterance is retrieved and recognised in the speaker's speaking comprehension system with reference to his or her long-term memory. Simultaneously, self-monitoring that detects the possible errors in the overt speech takes places (Kormos, 2006, 2011; Levelt, 1989; Xu, 2015; Yahya, 2019).

Monitoring

As delineated earlier, in a speaker's speech production, three loops of monitoring occur to inspect the outcome of the speaking process. The first loop of monitoring concerns comparing the preverbal plan produced in the conceptualisation with the speaker's initial communicative intention in the real situations. As the speaker's speech intension is determined by the given task, this loop of monitoring is to examine whether task demands on the speaker is met via the speaker's planning of the content and the language used in his or her intended discourse. If not, the monitor in the conceputaliser will send an alarm to the speaker who will resort to the relevant problem-solving mechanism to modify, partially replace or even gives up the original preverbal plan. The aim of such an action is to make sure that the generated new plan meets the task demands.

The second loop of monitoring is to examine if there are errors in the phonetic plan or the internal speech produced in the formulator before articulation. For instance, whether there are erroneously selected words. Such monitoring is also called covert monitoring in comparison with the overt monitoring in the third loop which is in charge of the inspection of the final utterance in the speaker's comprehension system via an acoustic-phonetic processor or the audition. During the final loop of monitoring, once errors are perceived, the speaker's monitoring system will issue a signal which will initiate a new round of speech production (Bygate, 1987, 2011; Kormos, 2006, 2011; Levelt, 1989; Xu, 2015; Yahya, 2019).

3.4.2.3 Problem-solving mechanisms

In L2 speech production, as a speaker's L2 knowledge may not be complete; it is unavoidable that the speaker will encounter problems (Dörnyei & Kormos, 1998). According to Dörnyei and Scott (1997), Dörnyei and Kormos (1998), and Komors (2006, 2011), there are four types of problems a speaker may have in L2 speech production: Resource deficit, time pressure, perceived deficiencies in the speaker's language output and the perceived deficiencies in the interlocutor's performance. The fourth type of problem is related to L2 speakers in conversation beyond this study (Kormos, 2006, 2011), and hence was not discussed.

Resource deficit refers to L2 speakers' knowledge gap caused by their L2 competence deficit which prevents them to verbalise their intended messages. Such a problem will occur when the speakers cannot retrieve the knowledge required by the input or the task demands from their long-term memory. It relates to three problem-solving mechanisms that take place in conceptualisation and formulation: Lexical problem-solving mechanism, grammatical problem-solving mechanism, and phonological problem-solving mechanism. As their names suggest, the three problem-solving mechanisms are associated with the speakers' knowledge gap in lexical items, grammar and the phonological forms of words. For problems as such, solutions include abandoning or changing the original speech plan, keeping the macro-plan unchanged while modifying the preverbal message, grammatical substitution, and phonological substitution. As L2 speakers frequently need more time to plan and process their L2 speech than that would be naturally available in a fluent communication setting, problems due to time pressure are unavoidable. Solutions to these problems are pauses and repetitions, such as the use of gap fillers. The third type of the problem is the incorrectness or inappropriateness of the speakers' utterance, and relevant solutions are self-repair, self-appraising and self-editing (Dörnyei & Kormos, 1998; Komors, 2006, 2011). It is evident that all the problem-solving mechanisms involved in L2 speech production are consistent with the working definitions of the metacognitive problem-solving strategy under investigation (see Table 3.1).

3.4.2.4 Attentional control

Kormos (2006, 2011) postulated that in a speaker' L2 speech production, conceptualisation, formulation and monitoring are subject to the speaker's conscious attentional control. Since an individual's attention resources are limited, the three processes will compete with one another for the attention available, and how the limited attention is allocated among the three processes will be influenced by task demands which indicates the cognitive complexity of a given task or task complexity. According to Kormos (2006, 2011), the competition among the three stages for attention supports Robinson's (2015) task complexity model along the dimension of resource-dispersing (Lambert et al., 2017; Kormos, 2011), one of the research foci of this study (see Section 3.3.4)

3.4.3 Framing the study in Kormos' (2011) model

Kormos' (2011) Bilingual Speech Production Model reveals the important role of strategic competence (planning, problem-solving and monitoring) and task complexity (attentional control) in L2 speech production (Skehan, 2016, 2018; Kormos, 2006,2011). It therefore justifies the establishment of the correspondences between this L2 speech model and Bachman and Palmers' (2010) framework of non-reciprocal language use and between the model and Robinson's (2015) Triadic Componential Framework. The correspondences make it reasonable to frame the current study underpinned by Bachman and Palmer's framework on assessment in the L2 speech model. Henceforth, the perceived theory on the relationships among the Chinese EFL speakers' strategic competence, task complexity and their performance in the specific context of the integrated speaking test can be constructed as delineated below.

When the input or an integrated speaking test task is given to the Chinese EFL speakers, they will first analyse the task demands (planning time, task type, prior knowledge and steps involved) to set their goals of speech production. Based on the analysis, the Chinese speakers retrieve relevant concept information from their long-term memory to plan the content information and the order of the information for the intended messages. After the concept is specified, the speakers will then think about the appropriateness of the language used in accordance with task type, which indicates the speakers' speech acts or their communicational goals. In this conceptualisation stage, the speakers may retrieve their prior knowledge stored in the episodic memory, and a preverbal plan will be produced accordingly. Before the plan enters the second processing stage of formulation, the monitor situated in these Chinese EFL speakers' conceptualisers will check if the preverbal plan has met the demand of the integrated speaking test task in terms of the content and the choice of language. If the task demands have not been met, a signal will be issued by the monitor, and the speakers will turn to their strategic competence, in particular, the problem-solving strategy for modifications and corrections until the preverbal plan generated in the conceptualiser meets the demand of the integrated speaking test task. Simultaneously, the pressure caused by the limited planning time offered by the test task may also trigger the speakers' use of problem-solving strategies such as using gap fillers, self-repetition and even a pause to help them gain more time for handling the speaking task.

During the linguistic encoding of the preverbal plan in the formulator, the Chinese EFL speakers will search appropriate lemmas with relevant grammatical and phonologic features in their mental lexicon to express the concepts or the content of the intended messages reflected in the preverbal plan. The processing outcome in the formulator is that the preverbal message is processed into concrete and meaningful words, phrases and sentences for the intended messages as the speakers' internal speech. In the process, if the speakers are not able to successfully encode linguistically the preverbal plan due to their insufficient language competence, they will employ the metacognitive problem-solving strategies to compensate for such incompetence. For example, if they cannot find a word to convey the intended message in the preverbal plan in their mental lexicon, they may alter their original concept specification in the plan and accordingly they can retrieve alternative words or new lemmas to substitute the old lexical item.

In the meanwhile, the monitoring will detect the errors or the inappropriateness of the speakers' internal speech that will be articulated in the form of sound in the articulator. The overt speech produced after the articulation will go through the final stage of monitoring in accordance with the information retrieved in the speakers' speaking comprehension system. If the information suggests that the speakers' actual discourses do not meet the task demands of the integrated speaking test task, the monitor in the conceptualiser will issue an alarm signal to the Chinese EFL speakers. As a result, a new round of the mechanism undelaying their L2 speech production will be started, during which the learners' strategic competence will be activated again in the conceptualisation to modify, and to self-correct so that the demand of the integrated speaking test task can be met by their actual utterances.

It should be noted that during monitoring in the different stages of the speech production, the Chinese EFL learners are assumed to use evaluation in concert with monitoring (O'Mally & Chamot, 1990) in that without evaluation, these speakers are unlikely to execute their comparison between the preverbal plan generated in the conceptualisation and the intended messages to be encoded. Similarly, when these speakers use the monitoring strategy to check the internal speech and the overt speech, they have to use evaluation; otherwise, they are not able to judge whether or not their actual utterance are consistent with the task demands of the integrated speaking test task (Purpura, 1999).

In summary, the variation in task complexity is assumed to cause modifications occurring in conceptualisation which is closely related to one's strategic competence (Skehan, 2016, 2018). The variance in task demands may also generate more problems in speech production, for which the Chinese speakers are expected to use their strategic competence for solutions more frequently and actively.

From the perspective of attentional allocation, since the Chinese EFL learners' attentional resources are limited in the speaking process, they have to distribute their attention among conceptualisation, formulation and monitoring (Kormos, 2006, 2011). When task demands are increased, the Chinese speakers are expected to allocate increasing attention to analyse the task characteristics and to plan the concept specification and the language choice in the conceptualisation. As a result, a more complex preverbal plan may be generated. To encode the plan with increased complexity from the perspective of linguistics, these speakers are very likely to invest more attention in retrieving the appropriate lemmas with relevant syntactic and phonetic characteristics in line with the plan. In consequence, when the speakers consciously increase the amount of their attention to conceptualisation and formulation, the attentional resources controlled by these speakers for monitoring will be reduced (Kormos, 2006, 2016). The attentional reduction in monitoring indicates that more errors may be undetected in various stages of the speech production, including the speaker's final speech. Hence, the quality of the speakers' performance will be negatively affected.

To summarise, when the Chinese EFL learners are given the four integrated speaking test tasks with varying degrees of complexity, the learners' respective attention on the conceptualisation and formulation will change accordingly. Such changes suggest that the speakers will allocate a varying amount of their attention to handle the conceptual meanings and the linguistic forms required to complete the integrated speaking tasks using their strategic competence. As the attention accessible to these learners is limited, they have to adjust the amount of their attention available for metacognitive monitoring. Since monitoring plays a part as the quality inspector in the speech production, any modifications on the speaker's attention on monitoring may affect the speakers' speaking performance.

The above delineation of framing this study in Kormos' (2011) model indicates independent correlations between strategic competence and task complexity, and between task complexity and speaking performance. Moreover, when the three variables are taken into consideration interactively, it can be seen that changes in task complexity will stimulate the variability in the Chinese EFL learners' use of strategic competence, which will further elicit changes in their speaking performance.

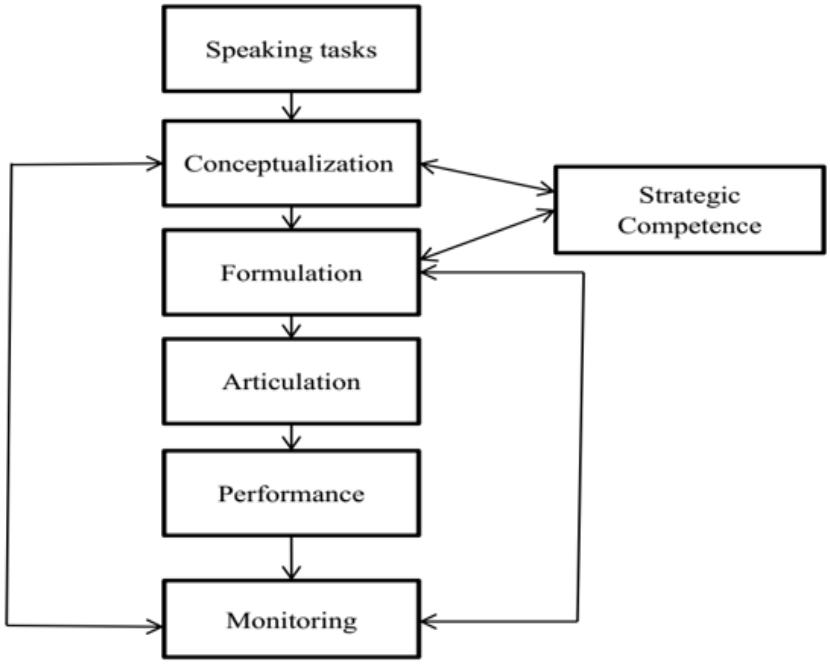

Figure 3.6 *Kormos' (2011) Bilingual Speech Production Model Guiding This Study*

In addition, the negative and direct correlations between task complexity and speaking performance can be justified. However, as monitoring functions in an overt form as one operating area of the speakers' strategic competence and simultaneously in a covert form as one of the four indispensable stages of L2 speech production, the complexity makes it hard to perceive the direct correlations between strategic competence and task complexity and between strategic competence and speaking performance respectively. Nonetheless, due to the problem-solving role of the strategic competence, particularly the quality-inspecting role of metacognitive monitoring in L2 speech production, it is tenable to state that although strategic competence cannot directly influence speaking performance, it moderates the effects of task complexity on the Chinese EFL learners' performance in the speaking process. Furthermore, in the four subordinates of the

Chinese EFL learner's strategic competence or their use of the metacognitive strategies, monitoring demonstrates the most influential moderating impact on speaking performance. The moderating effect indicates the moderator role of monitoring as shown in Figure 3.1. Figure 3.6 displays the perceived theory on the relationships among the three research variables in L2 assessment within Kormos' (2011) Bilingual Speech Production Model.

3.4.4 Measurement

In L2 assessment, speaking performance is often measured via rating scales (O'Grady, 2018; Sato & McNamara, 2019; Skehan, 2018). Rating scales are also known as scoring rubrics (Davis, 2018), and typically, a rating scale consists of a series of hierarchical levels, and each level presents a verbal descriptor illustrating a specific aspect of language proficiency. When these levels are combined, they constitute the operational language proficiency that test developers intend to assess on a test-taker. With reference to the descriptors, a rater can match test-takers' speaking performance to a specific level on the scale and reach a score (Davis, 2018; Fulcher, 2012a, 2018; Luoma, 2004). It explains why Ellis et al. (2019) stated that "testing typically ends with a score, and it is assumed that the score will reflect what we want it to reflect" (p. 246).

There are two types of widely applied rating scales: The holistic and analytical scales which correspond to the holistic scoring and the analytic scoring, respectively. A holistic scale permits raters to give a test-taker a single score based on their overall impression of the test-taker's speech performance quality. The general impression is in line with the verbal descriptors in a holistic scale which directs the raters' attention to various aspects of the test-taker's performance that are combined into a single score in the end. This performance measurement is holistic scoring which is regarded as timesaving because of its simplicity. In contrast, an analytic scale requires raters to judge a test-taker's speaking performance from various dimensions including grammar, structure, fluency and accuracy with separate scores being given for each of the dimensions. Compared with holistic scores, analytics scores are believed to provide a more detailed measurement of the test takers' performance, but they demand more time. However, for some raters, they may not make a clear distinction regarding different dimensions of a test-taker's speaking performance, and as a consequence, they are likely to award same

scores across the dimensions. This problem is called the halo effect, which turns the intended analytic scoring into holistic scoring (Barkaoui, 2010a, 2010b; Davis, 2018; Jamieson & Poonpon, 2013; Namaziandost & Ahmadi, 2019; Skehan, 2018).

Although current literature is replete with arguments for and against the two different scoring methods, as the integrated speaking test tasks are borrowed directly without any alternations from the TOEFL speaking section as noted earlier, it is natural that the holistic rubrics for the test was adopted. Accordingly, a holistic scoring was employed to measure the Chinese EFL learners' speaking performance. However, since the holistic scoring also involves a separate judgment of a test-taker's performance in line with multiple criteria on the rubrics, Jamieson and Poonpon (2013) commented that strictly speaking, TOEFL speaking scoring is a holistic scoring, but less strictly speaking, it involves both the holistic and the analytic scoring.

3.5 Prior Empirical Studies

To review prior empirical studies on the relationships among the Chinese EFL learners' strategic competence or their use of metacognitive strategies, task complexity and these learners' speaking test performance, a review of literature on the interrelationships between each two of the three variables is necessary as reflected in the research questions (see Section 1.6.3). In addition, given the relatively small amount of studies on L2 speaking assessment, my literature review was conducted typically from the perspectives of reading, writing, listening and speaking involved in the more recent studies with special focus on speaking.

3.5.1 Strategic competence and performance in L2 assessment

Empirical studies on metacognitive strategy use in the fields of metacognition and language learning strategies are "rigorous" (e.g., Amani, 2014; Chen, 2019; Rahimirad & Shams, 2014; Qin, 2018; Zhang & Qin, 2018). In contrast, the number of studies on strategic competence or metacognitive strategy use in L2 assessment, in particular, in the context of speaking is limited (Seong, 2014; Weir & O'Sullivan, 2011). Among them, Zhang (2014, 2017) investigated the relationship between 593 Chinese first-year college students' use of metacognitive strategies and their reading performance on the CET-4 (College English Test band four) reading subtest by using a multi-sample

structural equation modelling (SEM) approach. Data analysis of the participants' responses to a self-developed questionnaire revealed that planning, problem-solving (inference-making), evaluating, and monitoring significantly affected the Chinese EFL test-takers' reading performance in their lexico-grammatical reading ability, but these metacognitive strategies had a weak influence on their text comprehension ability. Indeed, Zhang's studies originated from Purpura (1997, 1998, 1999) and Phakiti (2003, 2008a, 2008b), and hence they share similarities in terms of research purposes and methods. However, the results of Purpura's studies indicated that EFL test-takers' use of metacognitive strategies (assessing and monitoring) did not relate to their reading performance. Likewise, with analysis of the data collected via questionnaires derived from Purpura's studies and interviews by the statistical means of multivariate analysis of variance (MANOVA), Phakiti (2003) examined the influence of 384 Thai EFL test-takers' strategic competence on their reading test performance. He found no significant correlations between the two variables. In contrast with Purpura's and Phakiti's studies, positive correlations between EFL learners' metacognitive strategies and their reading test performance were identified in a more recent study (Nourdad & Ajideh, 2019), though.

Regarding writing, Yang (2014) explored the interrelationship between test-takers' metacognitive strategy use and their performance in writing tests. Data were collected from 298 Taiwanese undergraduates in their writing test for summary essays and their self-reported strategy use. The study showed that planning and evaluating strategies served as administrative control over other strategies at varying degrees to improve summarization performance. In the context of listening, Pan and In'nami (2015) invited 170 Taiwanese university students to perform the Test of English for International Communication (TOEIC®) practice listening test, the sister products of TOEFL, and to respond the questionnaires to measure the participants' cognitive and metacognitive strategies. After exploratory factor analysis (EFA), repeated-measures MANOVA, and analysis of variance (ANOVA), the two researchers concluded that the participants' reported strategy use had a weak effect on their test scores, accounting for 7% of the total score variance, and in the three metacognitive strategies, monitoring and evaluation, in comparison with planning, had more influence on the participants' listening test scores. In terms of test-takers' overall language proficiency, Song and

Cheng (2006) have examined the correlations between metacognitive strategy use reported by 121 Chinese EFL learners and their performance on CET-4 through a questionnaire. Through statistical procedures such as factor analyses and multiple regression, they found that although Chinese EFL test-takers' use of metacognitive strategies had no direct effects on their performance, they reported using strategies such as problem-solving. Moreover, this strategy served as the best predictor of the participants' CET-4 test scores, accounting for 10% of the score variance. The two researchers, therefore, concluded that problem-solving played an important role in L2 assessment and merited further investigation.

In L2 speaking assessment, Fernandez (2018) examined the relationship between test-takers' strategy use and their performance in the third part of the IELTS speaking tests by means of stimulus recall. The participants were 12 EFL learners with diverse origins. The coding results suggested that there were no positive correlations between strategic competence and the participants' test performance. On the other hand, in a quantitative approach, Xu (2016) studied the relationship between Chinese EFL test-takers' use of speaking strategies and their IELTS speaking performance. A total of 93 Chinese postgraduate students were engaged and their strategy use was reported via a questionnaire. Data were analysed with correlation analysis, regression analysis and ANOVA, which revealed that Chinese EFL test-takers' performance was positively correlated with their use of planning, monitoring and evaluating. In a similar vein, Huang (2016) investigated 244 EFL learners' use of strategic competence and its correlations with their performance in a large-scale standardized English proficiency test in Taiwan. After sitting two sets of the test, the participants completed a survey inventory for data collection. Statistics testing methods of EFA and SEM were performed, and the findings showed that the participants' strategic competence directly influenced their test performance.

Overall, the research results on the effects of metacognitive strategy use on test performance are inconclusive (Nett et al., 2012; Dawadi, 2017; Song, 2005). Some indicated the positive relationship (e.g, Huang, 2016; Net et al., 2012; Nourdad & Ajideh, 2019; Rukthongand & Brunfaut, 2020; Sawaswati, 2017; Zhang, 2014), while the others evidenced the opposite trend (e.g., Barkaoui et al., 2013; Ghaemi & Ghaemi; 2011;

Fernandez, 2018; Pan & In'nami, 2015; Phakiti, 2003; Purpura, 1997,1998; Song, 2005). The mixed results suggest that the exploration of metacognitive strategies in L2 assessment is warranted (Seong, 2014; Song & Cheng, 2006).

3.5.2 Task complexity and performance in L2 speaking

Although there is a large percentage of empirical studies on the correlations between task complexity and speaking performance within Robinson's (2015) Triadic Componential Framework, they are mainly conducted in the non-testing situations or in the face-to-face context (Adams & Alwi, 2014). In addition, the speaking tasks adopted in these studies are often designed by the researchers for the manipulation of the assumed task complexity without validity (see Section 3.3.5), and the speaking performance is typically measured through discourse analysis, different from the scoring methods commonly employed in L2 speaking assessment.

Furthermore, these studies demonstrate two commonalities. The first commonality is that they are primarily operationalised dichotomously (e.g., simple versus complex), concentrating on the resource-directing dimension (Adams & Alwi, 2014; Tajeddin & Bahador, 2012). Regarding the resource-dispersing dimension, compared with studies into resource-directing variables, those investigated variables along the dimension of resource-dispersing are limited in numbers and among them most primarily focus on planning time (Awwad, 2019). Yuan and Ellis (2003), for example, examined the effect of tasks on oral production under the conditions of no planning, pre-task planning and online planning with a narrative task. Their study revealed that pre-task planning task condition increased lexical variety, and complexity in learners' oral performance, but had no significant effects on accuracy, whereas online planning task condition without pre-task planning improved grammatical accuracy over lexical variety. In terms of task type, several studies have revealed that task type affects task-takers' oral production under the conditions of varying planning time (Foster & Skehan, 1996; Neary-Sundquist, 2008; Skehan, 2009; Gilabert, Barón & Llanes 2009). A more recent study on this topic was done by Rashvand Semiyari and Ahangari (2019), in which 40 upper intermediate Iranian EFL learners performed four different types of speaking test tasks (explaining, problem-solving, storytelling, and picture-describing). Data analysis through one-way repeated measures ANOVAs, paired sample t-tests and Pearson

Product Moment did not suggest any considerable effects of task types on these Iranian EFL learners' speaking performance.

Another commonality is that researchers tend to manipulate one or two variables to establish the variance in task complexity, and hence studies on the simultaneously manipulation of the variables along the two dimensions in L2 assessment conditions are comparatively insufficient (Awwad, 2019). For instance, Gilabert (2005, 2007) established four levels of task complexity by simultaneously manipulating two variables: The pre-task planning time and the degree of displaced, past time reference (or +/-Here-and–No reference). His study revealed that pre-planning time had direct impacts on L2 learners' fluency, and lexical complexity; whist the manipulation of +/-Here-and–Now affected the participants' fluency and accuracy in their speaking performance. Likewise, using a repeated measures design, Levkina (2008) conducted a study on 14 Spanish and Russian students with an upper-intermediate level of English to explore the impacts of simultaneously manipulating task complexity along +/-planning time and +/- few elements on their L2 speech production. The results indicated that planning time had positive impacts on participants' lexical and structural complexity, but it did not affect fluency or accuracy significantly. Later, Levkina and Gilabert (2012) extended Levkina's (2008) research and found the synergetic impacts of changing task complexity variables on task performance.

Due to the above-mentioned commonalities, little is known about how task variables along the task complexity dimensions predict task performance when they are manipulated simultaneously in the authentic L2 speaking assessment context. Similarly, the literature on the interactions within these task factors when manipulated simultaneously is not available. As such, a study like this present one is justified.

3.5.3 Metacognitive strategies and task complexity in L2 assessment

Test-takers' metacognitive strategy use varies across tasks, which has been found by some researchers. Chou (2013), for example, has studied the strategic competence of 92 Chinese EFL learners in response to two different reading test tasks (general vs subject-specific). Analysis of the data collected from the post-testing questionnaires, and interviews suggested that participants used similar strategies across the two reading test

tasks. When the Chinese EFL readers failed to comprehend the reading texts given, they adopted their monitoring to solve the problem. In writing, Barkaoui (2015) has researched L2 test-takers' writing activities in handling two different writing tasks of the TOEFL iBT ® writing test: The Independent writing task and the integrated writing task. Participants (N = 22) responded to the two tasks before they reported their use of strategic competence via stimulated recalls. Coding results indicated that the participants' writing activities reflected their active use of the planning (local planning), problem-solving (revising language), monitoring (taking notes), and evaluating (evaluating language), which varied significantly across the test tasks. In Pan and In'nami's (2015) study on listening tests, (see 3.5.1 for a more detailed introduction to the study) planning strategies were mostly used in easier tasks, and such strategies had the highest frequency across task types, suggesting the importance of planning in listening tests. In addition, when task complexity increased, the participants were more likely to use planning, monitoring and evaluating.

Youn and Bi (2019) explicated the reported metacognitive strategy use across speaking assessment tasks. In their study, ESL students (N=30) completed four speaking tasks that differed in task complexity: A monologue speaking task about giving feedback to one's classmate's writing; two role-play tasks with one asking the test-takers to request a recommendation letter of a professor and the other one regarding refusing a professor's request on class change; and a role-play task on negotiating over the time and mode change related to a discussion activity. With retrospective reports, the participants' strategy use was elicited, transcribed and coded. The result showed that evaluating one's own performance and assessing task-related situations were the most frequently reported strategies across task types, and the use of metacognitive strategies varied across tasks.

Taking a quantitative research design, Kaivanpanah et al. (2012) examined the effects of task complexity on 227 Iranian students' strategic competence through questionnaires. Three different types of speaking test tasks were used: Picture description, telling a joke and telling a story. The findings demonstrated that differences in task demands, lack of context, time constraints, and the presence of the interlocutor had a significant impact on the participants' use of problem-solving.

It should be stressed that in these studies, metacognitive strategies were investigated individually, and their interactive working mode (see Section 3.2.4) were beyond examination. In fact, as noted in Section 1. 3. 2, researchers have commonly examined individual metacognitive strategies, and the interactive characteristic of the construct has been given little attention. In this sense, to better understand how metacognitive strategies work in L2 speaking assessment, it is imperative to investigate both the independent and the interactive working modes of the construct.

3.5.4 EFL learners' metacognitive strategies, task complexity and their speaking performance

In L2 assessment, it has been agreed that strategic competence and test tasks impact test-takers' speaking performance either independently or interactively (Huang et al., 2018; Fernandez; 2018; Fulcher, 2014; Phakiti, 2016; see also Bachman & Palmer; 1996, 2010; Canale & Swain,1980; Cohen, 2014; Huang, 2013; Swain, 2009; Weir & O'Sullivan, 2011). Researchers in this domain also agree upon the role of strategic competence as a mediator in the interaction between test tasks and test-takers (Bachman & Palmer; 1996, 2010; Barkaoui et al. 2013; Ellis et al., 2019). Despite these agreements, empirical studies on how test-takers' metacognitive strategy use, task complexity and test-takers' speaking performance interact with one another in the authentic L2 speaking assessment have been lacking (Barkaoui et al., 2013; Huang, 2016).

Based upon previous researchers' comprehensive review of the empirical studies concerning strategic competence, task complexity and test performance (e.g., Huang, 2016; Seong, 2014; Yi, 2012), I am aware that there are only four studies that cover all the three variables in an interactive perspective as with the case of the current study (Barkaoui et al., 2013; Huang, 2013, Swain et al., 2009; Yi, 2012). The scantiness may be due to the well-acknowledged actuality: "Language assessment is a complex field and one that the most experienced and highly regarded experts remain challenged by" (Hughes & Reed, 2017, p. 87), and in this field, speaking is under-researched (Fulcher, 2015a, 2015b; O'Sullivan, 2012; O'Sullivan & Weir, 2011).

Among the four studies, in an exploratory approach, Swain et al. (2009) explored the differences in test-takers' strategic behaviours in processing the integrated and the independent tasks in TOEFL-based speaking tests, as well as the relationship between test-takers' strategic behaviours and their oral performance reflected by their test scores. This study was conducted on 14 graduate and 16 Chinese undergraduate students with the method of think-aloud. The results showed that the Chinese test-takers used a total of 49 different strategies when completing all the speaking test tasks. Metacognitive strategies, communication strategies, and cognitive strategies were proportionally the most frequently reported strategies. Three metacognitive strategies were unique to the independent speaking tasks, which were monitoring, evaluating performance, and making choices. In the integrated speaking tasks, Task 3 and Task 4 elicited the test-takers' use of evaluating. This strategy was also reported by the test-takers in performing Task 5 and Task 6. Overall, the integrated speaking test tasks elicited a wider variety of strategies than the independent speaking test tasks, but there were no significant differences in metacognitive strategy use within the four integrated speaking test tasks. As for the correlations between the test-takers' metacognitive strategy use and their test scores, no direct relationship was found. Moreover, the researchers also found that there was no statistically significant difference in the test-takers' reported strategy use across their language proficiency levels. To conclude, the researchers stated that there were no relationships between the total numbers of reported strategic behaviours, including the metacognitive strategies and the total test score across the whole section of TOEFL-based speaking test tasks. Despite such results, Swain and colleagues still emphasised that strategy use was critical in test-takers' task performance on the integrated speaking test tasks as a mediator between test tasks and test-takers. Later, Barkaoui et al. (2013) updated this study and arrived at similar findings.

Inspired by Swain et al. (2009), Yi (2012) took an additional step by modelling their study in two conditions: A Testing condition and an academic classroom condition. Speech samples of six Korean EFL university students were collected on TOEFL-based speaking test tasks. Data on their reported use of strategies under the two conditions were elicited by stimulated recall verbalisation and were coded into five categories: Approach, compensation, cognitive, metacognitive strategies and their feelings. Participants' speech performance was rated with reference to the TOEFL speaking test

rubrics. Data analysis disclosed that metacognitive strategies were used the most frequently under both conditions. Compared with their performance on the independent speaking test tasks, the participants used more strategies on the integrated speaking test tasks in the two conditions, indicating positive correlations between the participants' metacognitive strategy use and task complexity. About the association between the participants' use of the strategies including metacognitive strategies and their speaking performance, the study revealed a weak relationship between the two research variables with the total strategy use accounting for less than 5.5% of the variance in their speaking performance in the test condition and less than 3% in the academic classroom condition respectively.

In a same vein, Huang (2013) probed the relationships among EFL learners' strategic competence, task type, and their performance in the IELTS Speaking test in testing and non-testing conditions. A total of 40 Chinese EFL university students were engaged in the study. With data collected from the stimulus recall, and the use of MANOVA, these participants' reported use of strategies across the three parts of the speaking test were analysed and the relationships among strategy use, task types and their speaking performance were assessed. Results showed that evaluating one's performance was one of the top-five individual strategies that the participants employed across task types in both the testing and the non-testing contexts. A significant difference in evaluating and planning across task type was discovered. However, these metacognitive strategies did not significantly correlate with the participants' test scores across tasks and contexts.

According to Seong (2014), although the relationships among test-takers' strategic competence, test tasks and test performance were investigated comprehensively in the above four studies, some problems still exist. First, researchers typically employed the stimulus recall or think-aloud solely to collect data on a small sample. The small sampling places the generality of the research findings into question. Similarly, the validity of the findings is also questionable, as data born in one method should triangulate with that of another (e.g., Creswell & Creswell, 2018; Creswell & Guetterman, 2019). Second, in terms of data analysis, in the studies conducted by Barkaoui et al. (2013) and Swain et al. (2009), merely frequency counts were used to delve into the participants' strategy use without a deep in-depth investigation of "who

uses each strategy, why, where, when, and how" (Swain et al., 2009, p. 56). In Huang's research, though the two parts of the speaking tests (the speaking section of the IELTS has three parts in total) involved face-to-face interviews or reciprocal speaking tests, the researcher did not take into consideration the interlocutors factors possibly due to the violation of assumptions in MANOVA. In addition to the problems pointed out by Seong, another fundamental problem is that test tasks in the four studies were conceptualised as holistic entities rather than a set of task characteristics as advocated in L2 assessment (Bachman, 2002; Bachman & Palmer 2010). Furthermore, as discussed in Chapter One, metacognitive strategies in the studies were investigated in an independent manner and the interactive working mode of the variable was not examined.

Taken together, the under-research of L2 speaking assessment, the inconclusiveness of the research results and the problems with the four prior studies pertaining to the present study suggest a research lacuna. An investigation into the relationships among strategic competence, task complexity, and speaking performance within the framework of non-reciprocal language use via a mixed method research design on a larger sample size in this study is therefore expected to fill the research niche and complement the existing relevant literature as noted in Section 1.2.

Chapter Four

Research Design and Methodology

4.1 Chapter Overview

"Research is a process in which you engage in a small set of logical steps" (Creswell & Guetterman, 2019, p. 2). In this chapter, I presented the "logical steps" involved in this research by rationalising the research design and methodology underpinning these steps. The presentation begins with an overview of the research design followed by the selection of the research sites and participants. A detailed description of research methods including instrument development and validation, data collection and data analysis is then provided. The chapter ends with ethical considerations.

4.2 Research Design

Overall, a convergent mixed methods design was adopted. As one of the major types of the mixed methods design, the convergent mixed methods design combines the quantitative and qualitative approaches into one enquiry. A quantitative approach helps researchers verify theories by examining relationship among variables measured on instruments and analysed in statistical procedures. On the other hand, a qualitative approach enables researchers to investigate a social phenomenon from the perspective of individuals or groups with data involving participants' setting and data analysis built inductively from particular to general. With the strengths of the two approaches, the convergent mixed methods design makes feasible a comprehensive exploration of a phenomenon (Creswell & Creswell 2018; Creswell & Guetterman; Creswell & Poth, 2018).

In addition to the feasibility, the selection of a research design is typically in line with a researcher's assumption, the procedure of enquiry for a research objective, research methods for data collection, analysis and interpretation, and research questions (Creswell & Creswell 2018; Creswell & Guetterman, 2018; Creswell & Poth, 2018). With reference to these criteria, the convergent mix methods design was employed in the current study for the following reasons:

a) The research questions investigated the Chinese EFL learners' reported use of metacognitive strategies, their oral production performance, and the complexity of the integrated speaking test tasks. Data on metacognitive strategies was collected from questionnaires which were integrated with oral interviews. Data analysis accordingly involved not only statistical procedures but qualitative coding. The two types of datasets were converged to investigated the research questions.

b) Data on participants' perceptions of task complexity in the four integrated speaking test tasks was collected via self-rating scales, self-rating open questionnaires, and semi-structured interviews (see Section 3.3.6). The integration of the quantitative and qualitative instruments justified a convergent mixed-methods research design.

c) According to Hughes and Reed (2017), to research speaking, "multitude of ways" should be adopted, since the target of such research is "dynamic and socially grounded" spoken language (pp. 123-125). As integrated speaking test tasks were examined in the present study, a convergent mixed method design could provide "a multitude of ways" to study the Chinese EFL leaners' performance elicited by the integrated speaking test tasks. Literature on speaking performance also supports the use of a convergent mixed- methods design (e.g., Huang & Hung, 2013; Huang et al., 2016; Sun, 2016).

Within the convergent mix methods design, a one-way repeated measures design was employed to collect the data on the Chinese EFL learners' metacognitive strategy use, their perceptions of task complexity, and their oral scores on the four integrated speaking test tasks. Scholars explain that a one-way repeated measures design involves multiple measures of the same research variable conducted on the same participants either under different treatment conditions or over different time periods (Gilabert, 2005, 2007a, 2007b; Levkina, 2008; Rahimi, 2016; Weir et al., 2006). In line with the design, the Chinese EFL students answered one Metacognitive Strategy Inventory (MSI) and one self-rating scale after they finished one integrated speaking test task which represented one treatment condition. Such procedure repeated on the four speaking tasks for four

treatment conditions on task complexity as shown in Table 3.2. Two trained raters scored the students' speaking performance by means of holistic scoring with reference to the TOEFL rubrics for the integrated speaking test tasks (Huang et al., 2018; Huang & Hung, 2013; Luoma, 2004; Nett et al., 2012). Data collected were either compared or correlated for addressing research questions via multiple statistical procedures including: ANVOVA, MANOVA, Pearson Moment Product correlation, a multiple linear regression and Hierarchical Linear Model (HLM) (e.g., Barkaoui, 2010, 2013; Hair, 2019; Hair et al., 2014; Kline, 2016; Meissel, 2014; Mujis, 2011; Nett et al., 2012; Qin, 2003; Raudenbush & Bryk, 2002).

With regard to the qualitative data on the students' metacognitive strategy use, and on task complexity perceived by the students and the teachers, a phenomenological design was used. This design focuses on the commonality concerning the specific experience of a group of participants related to a particular phenomenon for the description of that phenomenon (Creswell & Guetterman, 2019; Creswell & Poth, 2018; Daniel, 2011). Data on the students were collected from a sub-set of the student participants on the Semi-Structured Interview Guide (SSIG), and were analysed on NVivo 12 (Windows version), while the teachers' data were gathered on the open-ended questionnaires and were analysed manually (Révész, 2014; Révész et al., 2016). In the convergent mixed methods design, the phenomenological design served two purposes: First, it facilitated a more in-depth examination of the Chinese EFL learners' metacognitive strategy use in the four treatment conditions, and of their perceptions of task complexity, thereby converging and aligning the quantitative data on the two variables (Creswell & Guetterman, 2019; Creswell & Poth, 2018; Harris & Brown, 2010). Second, it helped to assess task complexity from the perspectives of the students and the teachers. The assessment formed triangulation needed for the credibility of quantitative analysis of task complexity (Révész, 2014; Révész et al., 2016; Sasayama, 2016; Wilson, 2013).

Within the research design, the present study was composed of two phases: Phase one focused on the development, validation and piloting of the inventory via the exploratory factor analysis (EFA) and the confirmatory factor analysis (CFA), scoring training, equipment testing and the organization of the speaking tests; Phase two targeted at data

collection and analyses to answer the research questions. Table 4.1 presents an overview of the mixed-methods design.

Table 4.1 *Overview of the Mixed Methods Design*

Phase	Research Objectives	Instruments	Stages	Participants
Phase One	Instrument development, validation and piloting	MSI, self-rating scale; SSIG	Developing, validating and piloting MSI and SSIG; Testing equipments and piloting the organization of the testing sessions; Rater training and scoring piloting	496 Chinese EFL learners
Phase Two	Quantitative: Evaluating task complexity, examining the effects of task complexity on Chinese EFL learners' metacognitive strategy use; the effects of task complexity on their speaking performance; the effects of Chinese EFL learners' metacognitive strategy use on their speaking performance; the relationship among Chinese EFL learners' metacognitive strategy use, task complexity and their speaking performance;	MSI, self-rating scales, TOEFL integrated speaking test, and rubrics for TOEFL integrated speaking tests	Quantitative data collection and analyses;	120 Chinese EFL learners and two raters
	Qualitative: Students' perceptions of their metacognitive strategy use and task complexity; teachers' perceptions of task complexity	SSIG and open-ended questionnaire	Qualitative Data Collection and analyses	Five Chinese EFL teachers and eight Chinese EFL students

4.3 Description of and Decision on Research Sites

The research sites are two Chinese universities in a Northern city in the People's Republic of China. The decision on the two universities was based on the followings.

4.3.1 Sample availability

The two universities provided me with the access to the sample of the Chinese EFL learners. Additionally, the characteristics of the students from the two universities (see section 4.4.1) justified the selection of the two universities (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Creswell & Poth, 2018; Daniel, 2011).

4.3.2 Time and financial constraint

According to Walford (2001), researchers normally "settle for a research site to which they can easily gain convenient and ready access…" (p. 151) due to budget and time strain. I chose the two universities mainly because they were where I had worked as an English teacher for their international cooperation programmes. It was comparatively convenient for me to seek permission and cooperation from the relevant departments. In addition, the location of the two universities also rendered convenience, as they were not physically distant from my accommodation during the fieldwork, which not only reduced the cost and time on commuting, but increased research efficiency as well.

4.3.3 Familiarity

The abovementioned working experiences in the two universities familiarise me with the research sites. Such familiarity helped me avoid culture shock or disorientation, offered me the possibility of enhanced rapport and communication with participants, improved the chances to gauge the honesty and accuracy of responses, and increased the likelihood of more intimate details of their lives from the participants to me. From this perspective, the decision on the two universities facilitated the smooth procedure of the research for efficiency, and enhanced the validity and credibility of the research results (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Creswell & Poth, 2018).

4.4 Participants

This study involved 616 Chinese EFL students, five Chinese EFL teachers and two raters, among whom eight students engaged in the semi-structured interviews. The

sample size was in accordance with relevant literature statistically (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Creswell & Poth, 2018).

4.4.1 Decision on students

Almost all the students were enrolled either in the Faculty of Foreign Language Studies or the International Cooperation Programmes in the selected research sites. The students in the International Cooperation Programmes were in their finial year before studying abroad and were familiar with the international language testing systems, whilst those from the Faculty of Foreign Language Studies were in their last academic year before their internship related to English. Because of the English–related background, the students were enthusiastic about this study which, they believed, potentially benefited them in their language preparations for their future study or career. This enthusiasm contributed to the students' cooperation, which helped improve the accuracy of their responses, and hence the validity was enhanced (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Creswell & Poth, 2018; Daniel, 2011).

Additionally, the score range of the students on CET-4, an authoritative test for English language proficiency in China (Zhang, 2017), was from 425 points to 500 points. According to the official scoring interpretations of the test published by the National Education Examinations Authorities (2020), such a score range suggests that the students' language proficiency were at an upper-intermediate level, which enabled them to distinguish simple tasks from the complex ones (Rahimi, 2016). Consequently, the validity of their perceptions of task complexity in the four integrated speaking test tasks that require rather higher language proficiency was established (Kyle et al., 2016).

By means of convenience sampling, a total of 616 voluntary Chinese EFL learners were recruited. After the approval was granted by the University of Auckland and by relevant departments in the research sites, a recruiting campaign for volunteers was launched both on-line and off-line. In the on-line promotion, an advertisement on this study was posted on the forum on the official websites of the two universities with the assistance of the secretaries from the administration departments, since the forum is usually what students pay attention to. Moreover, posters about the study were posted on the bulletin board. To arouse the interest of more students, an off-line recruitment was organized

campus-wide simultaneously. Some English teachers were encouraged to promote the volunteer recruitment between intervals in their classroom instructions. In the meanwhile, I organised a series of promoting activities among the potential candidates, giving them a brief introduction to this study via PowerPoint slides.

The objective, potential benefits, the significance of the study, and the language requirements were stated clearly during the recruitment. Additionally, rules on protecting participants' privacy and human rights related to ethical issues were also introduced. In the recruitment, I purposefully controlled the number of the participants in light of the effective sample size required statistically. The participants in the semi-structured interviews were also selected on a voluntary basis with reference to their responses to the last question of the questionnaire which surveyed respondents' willingness to participate in interviews.

4.4.2 Description of students

Students were invited to measure task difficulty on one-way repeated measures ANOVA in the two phases, and their number reached to 600 after data cleaning, which was consistent with the sampling requirement (Pallant, 2016; Green & Salkind, 2010). Validating MSI in Phase One involved EFA and CFA, which had specific requirement on sample size. A total of 496 students participated, with 254 for EFA and 242 for CFA. The sample sizes were based on the statistical requirements of factor analyses. EFA and CFA require at least 100 and 200 samples respectively with two different sample groups (e.g., Hair, 2019; Hair et al., 2014; Kline, 2016; Qin, 2018; Teng et al., 2018; Zhang, 2017). The students on the EFA were aged from 20 to 21 years, among whom 36.32% were male while female students accounted for 63.68%. The students' characteristics related to their English learning experiences are presented in Table 4.2.

Table 4.2 *Characteristics of the Students in EFA*

English Learning Experience	Performance	Number	Percentage (%)
Learning Length (Years)	7-9	105	49.53
	10-12	75	35.38
	13-15	22	10.38
	Others	10	4.72
Tests Participated	CET-4	209	98.58
	CET-6	124	58.49
	BEC	2	0.94
	IELTS	1	0.47
	TOEFL	1	0.47
CET4 Scores	＜425points	3	1.42
	425-500 points	175	82.55
	500-550 Points	27	12.75
	>550 points	7	3.30

Note. BEC = Business English Certificate; CET-6 = College English Test band 6.

In the CFA, valid responses were generated from 242 participants, among whom 90 were male students, accounting for 38.96% of the sample and141were female students. The age of the participants ranged from 18 to 21 and their English learning experience varied considerably as shown by Table 4.3.

Table 4.3 *Characteristics of the Student in CFA*

English Learning Experience	Performance	Number	Percentage (%)
Learning Length (Years)	7-9	95	41.13%
	10-12	93	40.26%
	13-15	20	8.60%
	Others	23	9.90%
Tests Participated	CET-4	228	98.7%
	CET-6	101	42.72%
	BEC	3	1.3%
	IELTS	1	0.43%
	TOEFL	1	0.43%
CET4 Scores	＜425points	14	6.06%
	425-500 points	134	58.01%
	500-550 Points	58	25.11%
	>550 points	25	10.82

Note. BEC = Business English Certificate; CET-6 = College English Test band 6.

Phase Two involved 120 students, and the number was in accordance with the comprehensive consideration on the thumb-up rules of the statistical testing methods employed. Data screening left 104 cases with male students making up 34.62% (N= 36). The students were aged from 18 to 20 and their English learning experience varied as displayed in Table 4.4.

Table 4.4 *Characteristics of the Student in Phase Two*

English Learning Experience	Performance	Number	Percentage (%)
Learning Length (Years)	7-9	44	42.31%
	10-12	44	42.31%
	13-15	12	11.54%
	Others	4	3.85%
Tests Participated	CET-4	102	98.08%
	CET-6	65	62.85%
	BEC	1	0.96%
	IELTS	6	5.77%
	TOEFL	2	1.92%
CET4 Scores	＜425points	13	12.5%
	425-500 points	56	53.85%
	500-550 Points	22	21.15%
	>550 points	13	312.5%

Note. BEC= Business English Certificate; CET-6 = College English Test band 6.

After assumption testing, an indispensable step in statistical procedures (Hu & Plonsky, 2019), a valid sample size of 95 was obtained. With respect to the semi-structured interviews, the sample size was eight, which met the rule of thumb-requirement of a phenomenology design that usually expects 5-25 participants (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Daniel 2011). On average, the interviewees had 11 years of English learning experience, and half of them had sit the IELTS. Table 4.5 reveals the interviewees' background information.

Table 4.5 *Background Information of Interviewees*

ID	Gender	Age	Learning (Years)	English Test Participated
In1	Male	20-21	10-12	BEC
In2	Female	21-22	13-15	CET-6
In3	Female	20-21	10-12	CET-6
In4	Male	19-20	10-12	IELTS
In5	Male	18-19	13-15	IELTS
In6	Male	18-19	10-12	IELTS
In7	Male	21-22	7-9	IELTS
Int8	Male	18-19	13-15	OTHERS

Note. In=Interviewee.

4.4.3 Teachers

The number of teacher participants (N = 5) was in accordance with the literature on cognitive measurement on task complexity (e, g., Révész, 2014; Révész et al., 2016; Sasayama, 2015, 2016; Xu, Zhang, & Gaffney, 2021). The teachers were native speakers of Chinese aged from 30 to 41, and all of them had more than 10 years of English teaching experience with a master's degree in English. Their background information is presented in Table 4.6.

Table 4.6 *Background Information of Teacher Participants*

ID	Gender	Age	Teaching Experience (Years)	Education Background
T1	Female	38	13-15	Master
T2	Female	37	13-15	Master
T3	Female	37	10-12	Master
T4	Male	36	13-15	Master
T5	Male	41	17	Master

Note. T=Teacher.

4.4.4 Raters

Raters were two Chinese EFL teachers with more than seven years of English teaching experience and a master's degree of English literature. Both were actively engaged in the cultural exchange programmes in the research sites and were very familiar with international language testing systems. One of the raters was a male and the other one was a female.

4.5 Instruments

As discussed in Chapter Three, the Metacognitive Strategy Inventory, the Semi-structured Interview Guide, the self-rating scales, four TOEFL integrated speaking test tasks and the test rubrics for the TOEFL integrated speaking test were used to elicit the students' use of metacognitive strategies, their perceptions of task complexity, and their speaking performance. Similarly, the open-end questionnaires were adopted for the Chinese EFL teachers' evaluations of task complexity as expert judgment to triangulate the students' perceptions.

4.5.1 Descriptions

The following subsections present the descriptions of the above instruments from two aspects: Students and teachers.

4.5.1.1 Instruments for students

The Metacognitive strategy Inventory (MSI)

The students' metacognitive strategy use was elicited by a newly developed MSI, which was developed on the four questionnaires reviewed in Section 3.2.6. However, as none of the four questionnaires is completely consistent with the research purpose of the current study, modifications on them are unavoidable. Thus, I integrated the four questionnaires in accordance with the definitions and the taxonomies of metacognitive strategies defined in this study (see Table 3.1), and devised the Metacognitive Strategy Inventory.

The metacognitive strategies elicited by the questionnaire were classified into four types: Planning, problem-solving, monitoring and evaluating, which were manifested by 23 items. A sample item on planning was "I knew what the task questions required

me to do". A sample item concerning problem-solving was "I drew on my background knowledge to complete the task". Items such as "I knew when I should complete a task more quickly" were used to examine the participants' use of monitoring. Similarly, "I evaluated whether my intended plans worked effectively" was one of the item scales that investigated the use of metacognitive evaluating strategy. Five structured questions on the students' background information such as age, gender, and their EFL learning experience were also included in the questionnaire. The closing question of the questionnaire was to recruit interviewees, which asked those who had the interest in the interviews to provide their email address for further contact.

A seven-point Likert scale was used for each item: 0 (not at all true of me), 1 (not true of me), 2 (slightly not true of me), 3 (neutral), 4 (slightly true of me), 5 (true of me), and 6 (very true of me) (Devellis, 2012; Teng, 2016). Though the questionnaire was developed in English, each item was operationalized as a written statement in Chinese, the students' native language, to eradicate possible misunderstandings and hence to enhance the reliability of the instrument (e.g., Dörnyei, 2010, Qin, 2018; Teng, 2016).

The Semi-structured Interview Guide (SSIG)

A Semi-structured Interview Guide (SSIG) was used to measure the Chinese EFL learner's metacognitive strategy use and their perceptions of task difficulty. The SSIG is comprised of 10 prompts or questions in two groups. As noted in Chapter Three, the first group concerns the metacognitive strategies reported by the students, which were developed based on Zhang's (1990) Scheme for Eliciting Subjects' Metacognitive Knowledge. A sample question is "Could you tell me what you did in the planning time for each task?" The second group regards task complexity and questions in this group were borrowed from Tavakoli's (2009) semi-structured interviews guide on task difficulty. Questions in this group include "You have done six speaking tasks, and which one do you think is the easiest one and which one do you think is the most difficult one? And why?" Prompts in the first group are highly similar to the items in the MSI and those in the second group on task complexity factors are generally within Robinson's (2015) Triadic Componential Framework for data alignment (Harris & Brown, 2010). The background information on the interviewees is also included in the guide. Figure 4.1 demonstrates the scheme underpinning the guide.

Research Objectives	Constructs	Questions Asked of Student Participants
Metacognitive Strategies	Planning	Could you tell me what you did in the planning time for each task?
		Could you tell me whether you had a particular goal before you started your task today?
	Monitoring	Could you tell me whether you completely understood the reading and listening materials in the tests?
	Problem-solving	Could you tell me what you did to solve the problems in the tests if you had?
	Evaluating	Could you tell me whether you had evaluated your performance in the test after the task was done?
Task Complexity	Overall Assessment	You have done 4 tasks in total, and which one do you think is the easiest one and which one do you think is the most difficult one? Why
	Planning Time	Do you think it would be easier for you to complete the task when you were given 30s to prepare compared with 20s given? Why?
	Steps Involved	Do you think it would be easier for you to listen before speaking than to read and listen before speaking? Why?
	Task Types	Do you think it would be easier for you to explain a concept than to providing a solution? Why?
	Prior Knowledge	Do you think it would be easier for you to speak on a topic related to campus life than to speak on a topic regarding an academic lecture? Why?

Figure 4.1 *The Scheme Underpinning the Semi-structured Interview Guide*

The Self-rating scale

The self-rating scale on task complexity was borrowed directly from Révész et al. (2016). It has two items: One relates to the students' evaluations of their mental efforts in performing the test tasks, and the other regards their perceptions of task difficulty. The two items were rated on a nine-point Likert scale with 1 suggesting that the task is

not difficult at all, whilst 9 indicating that the task is extremely difficult. As the native language of the students is Chinese, the scale was translated from the original language of English to Chinese. Figure 4.2 illustrate the original self-rating scale.

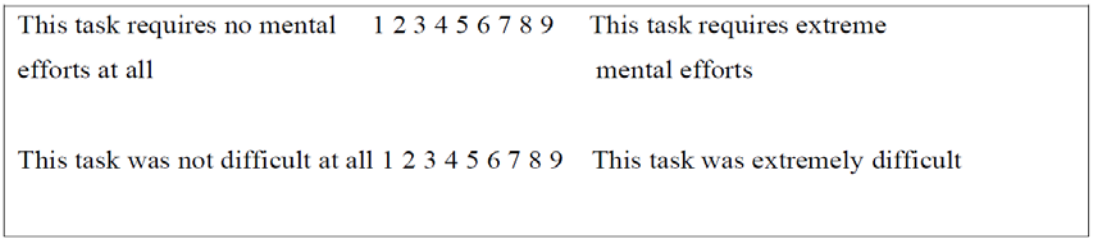

Figure 4.2 *Self-rating Scale*

Source: Measuring Cognitive Task Demands Using Dual-Task Methodology, Subjective Self-Raring, and Expert Judgements: A Validation Study, by A. Révész, M. Michel, & R. Gilabert, 2016, *Studies in Second Language Acquisition*,*38*(4), p.716. Copyright 2015 by Cambridge University Press.

Given the fact that there were only two items on the scale, it was presented after the students' background information in the MSI. Consequently, the students did not bother to respond twice in terms of their background information if the scale was presented independently, which made the instrument user-friendly (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Dörnyei, 2010).

TOEFL-based integrated speaking tasks

As one of the most popular high-stakes tests worldwide, TOEFL, or The Test of English as a Foreign Language, is designed to assess the language proficiency of EFL learners (Hughes, 2017; Hughes & Reed, 2017). It is an internet-based test focusing on academic English. The development of the test has brought about many studies to prove its validity and reliability. As a result, the validity and the reliability of the instrument have been established (e.g., Brooks & Swain, 2014; Brown & Ducasse, 2019; Chapelle, 2011; Farnsworth, 2013; Kyle et al., 2016; Luoma, 2004; Norris, 2008).

The TOEFL-based integrated speaking test tasks played a dual role: An indicator of task complexity (see Section 3.3) and an instrument to elicit the students' test performance. To ensure that the students could distinguish task complexity across the four tasks, the selection of the speaking test tasks was in line with the students' language proficiency

(Crossley & Kim, 2019). In addition, the contents of the speaking test tasks were characterised by "cultural neutrality, religious neutrality, and low controversy-provoking possibility" (Huang & Hung, 2013, p. 250). As a result, four TOEFL-based integrated speaking test tasks (Task 3, Task 4, Task 5 and Task 6) were chosen from the second set of the integrated speaking tasks from TOEFL practice online data (TPO2) without any changes for authenticity, validity and reliability (Chen, 2007; Farnsworth, 2013; Kyle et al., 2016).

TOEFL practice online tests are official practice tests that feature real past test questions and aim at allowing learners to experience taking the real TOEFL iBt test (ETS, 2021b). This ensures the authenticity, validity and reliability of the four tasks adopted in my study. It has to be stressed out that the four speaking tasks come from the old version of TOEFL iBt integrated speaking test which experienced a new round of the reform in late 2019.

The input format reflected by the four taks has been shown in Table 3.2, and the contents of the tasks were presented as the followings:

Task 3 was composed of a reading passage on a university's new plan on changing a shuttle route which was followed by a discussion between two university students on the plan in the listening section. After that, the student participants were required to state one of the speakers' opinion on the new change.

Task 4 provided a reading passage on a psychological concept: Audience effect. In the follwoing listening material, a lecture on this topic was delivered, and students were asked to use the examples given in the listening to explain the concept in the reading material.

Task 5 involved a conversation between a professor and a female student on a time conflict. To solve the conflict, the professor offerd the female student two possible solutions, with neither sounding satifisfactory to her. The student participants were required to recommend one specific solutions to the conflict and give the reasons why they believed such solution might work.

Task 6 was a lecture on two definitions of money in the listening section. The broad definition refered to both coins and bills and the barter system. The narrow definition indicated the leagal tender or whatever was accepted as payment such as coins and bills in a society. The students were asked to explain the two forms of money with the examples used by the professor in the lecture.

TOEFL integrated speaking test rubrics

The TOEFL integrated speaking test rubrics was developed by ETS in 2008, and it consist of four criteria: Delivery (fluency, clarity of ideas, and pronunciation), language use (grammatical accuracy and use of vocabulary), topic development (cohesion and progression of ideas) and general description (Huang et al., 2018; Putlack & Link, 2009).

4.5.1.2 The instrument for teachers

An open-ended self-rating questionnaire on task difficulty was administered on teacher participants for expert judgement as reviewed in Section 3.3.6, and it had two parts : The self-rating questionnaire and the participants' background information. The questionnaire was also borrowed directly from Révész et al. (2016), and its format and wording were almost the same with that for students. The only difference was that the scale for teachers had an open-ended question to examine their perceptions of task complexity in the four integrated speaking test tasks. The scale was presented in English after the teachers' language proficiency was taken into account. Background information on the teachers' English learning and teaching experiences were included at the beginning of the scale. Figure 4.3 illustrates the open-ended scale for the teachers, and Appendix 3 shows the full version of the questionnaire.

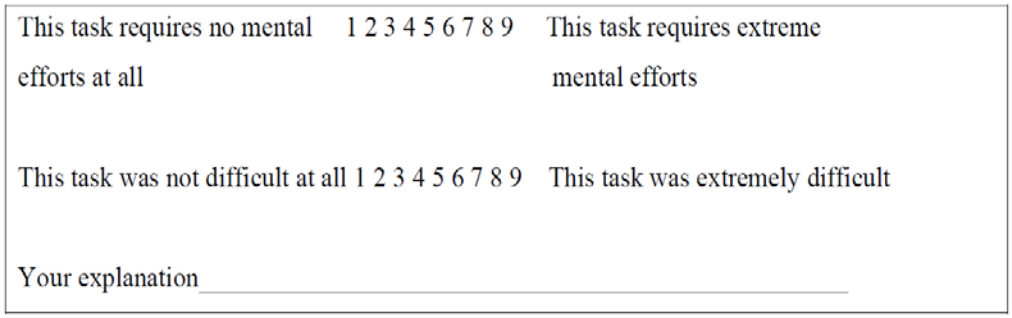

Figure 4.3 *Open-ended Self-rating Scale*

Source: Measuring Cognitive Task Demands Using Dual-Task Methodology, Subjective Self-Raring, and Expert Judgements: A Validation Study, by A. Révész, M. Michel, and R. Gilabert, 2016, *Studies in Second Language Acquisition*,*38*(4), p.716. Copyright 2015 by Cambridge University Press.

4.5.2 Validation and piloting

Since the MSI and the SSIG were newly developed, they were validated before actual use. In the similar vein, as the self-rating scale was translated from English to Chinese, its validation was imperative. Moreover, as the study involved a large sample size, instrument piloting was equally critical. Through validation and piloting, the reliability, validity and practicality of the instruments were achieved (Creswell & Guetterman, 2019; Creswell & Poth, 2018; Dörnyei 2010).

4.5.2.1 The Metacognitive Strategy Inventory

Validating the questionnaire relates to its face validity, content validity and construct validity (Hair, 2019; Hair et al., 2014; Muijs, 2011; Pallant, 2016). For face and content validity, I consulted four PhD students majoring in linguistics from the University of Auckland on the layout, wording, redundancy, and logic consistency of the questionnaire. As a result, one item that caused misunderstanding was removed. After this, two linguistic professors with the background of English and Chinese from Shanghai International Studies University (SISU) were invited to examine the translation of the questionnaire from original English to Chinese. They scrutinised the items in redundancy, sequencing, clarity, readability, and comprehensibility (Muijs, 2011). Based on their opinions, all the possible instructions, interpretations, and the scale items that might lead to misunderstanding were revised (Chapelle et al, 2011). Modifications were made in item wording, and one new item was added. The modified

questionnaire was then piloted on 22 potential participants to evaluate the wording, the structure and the clarity of the items for the readability and the understandability of the instrument in its actual users.

After piloting, the questionnaire was subject to construct validity comprised of two factor analyses, which was followed by reliability examination with reference to the Cronbach's alpha coefficient. The first factor analysis extracted 28 items under four factors representing four metacognitive strategies for further examination via the second factor analysis (Byrne, 2016; Kline, 2016; Pallant, 2016; Teng, 2016; Sun, 2016). The final validated questionnaire was composed of 23 scale items (see Appendix 1).

4.5.2.2 The Semi-structured Interview Guide

The interview guide was presented in Chinese translated from English to avoid understanding barriers. The translation was based on my consultation with one Chinese PhD student who has linguistic background from the University of Auckland. The two professors from SISU were again invited to back-translate the guide for the accuracy of the translation. After back-translation, the guide was piloted on two student participants for its comprehensibility. The length and the wording of some questions were revised after the piloting (Creswell & Guetterman, 2019; Creswell & Poth, 2018; Dörnyei 2010) (see Appendix 2).

4.5.2.3 Self –rating scales

Given the fact that the original self-rating scale was presented in English, but its actual users were Chinese EFL learners, I translated it into Chinese after consulting the two linguistic professors from SISU for validation (Creswell & Guetterman, 2019; Creswell & Poth, 2018). Subsequently, it was piloted on some student participants and no modifications were made because the scale was completely understood by the students.

4.5.2.4 TOEFL integrated speaking test tasks

The four integrated speaking test tasks were borrowed directly from TOEFL, a validated internationally acknowledged test with good reliability as noted previously. As the test was computer-delivered and some students might not familiar with the test format, piloting the test was necessary. Yet, the focus was not on the test per se, rather, the

piloting work concentrated on testing administration and testing equipments. Through the piloting, the students familarised themselves with how to use headphones and TPO softwares for the smooth delivery of the speaking test tasks and the successful storage of their speech samples for speaking performance. It should be noted that in validating and piloting the questionnire and the self-rating scale, the students had to sit the integrated speaking test tasks before the two instruments were administered on them. Henceforth, piloting the test tasks took place simultaneously during validating and piloting the questionaire and the self-rating scale, which increased research efficiency.

4.5.2.5 TOEFL integrated speaking test rubrics

Piloting the rubrics was essentially for rater training and piloting scoring to enhance intra-rater reliability and inter-rater reliability (Gwet, 2014; Hallgren, 2012). Rater training began with the two raters' familiarity with the rubrics by learning relevant materials presented on the official website of ETS. After this, they performed the four integrated speaking test tasks in person for acquaintance with the prompt. In the following scoring piloting, with reference to the rubrics, the raters first worked together to assess two actual performances from the students and reached agreed-upon scores for each performance. Then they assessed another set of two actual performances from the students independently, and when they finished, they explained why they gave a certain score to a specific performance. Such explanation was to improve the consistency between raters (Huang et al., 2013; Huang et al., 2018).

4.6 Data Collection

Data collection began in the early April of 2018 and ended in the late June of 2018, spanning half of the second academic term. The collected data were then input into relevant software packages for analysis.

4.6.1 Phase One

In Phase One, data collection was conducted for instrument validation on the Metacognitive Strategies Inventory, and for the Chinese EFL learners' perceptions of task complexity via the self-rating scale.

During instrument validation, the first set of student samples were invited to answer the

first draft of the questionnaire following the four integrated speaking tasks. The students performed the speaking tasks in multimedia laboratories, with each computer being given a number indicating the group number of the students. The number and the task sequence (see Table 4.10) created the students' codes in which the task sequence revealed the identification of their judgments on task difficulty and their oral performance on each of the speaking task. Before task performance, the students filled the Consent Form or CF (see Appendix 3) and the Participant Information Sheet or PIS (see Appendix 4) under the instructions of the researcher and the research assistants. Questions from the students related to the present study were solved. In answering the questionnaire, the students were asked to tick a number reflecting the relevant strategies they used in performing the integrated speaking test tasks. In responding to the self-rating scale, they selected a number from the range of 1 to 9 to represent their evaluations of task difficulty. Data collected from the first group of the students were used for the EFA. Another different sample of students answered the revised questionnaire generated in the EFA after they completed the integrated speaking test tasks. Data collected from this group was used for the CFA, following the same procedure as in the EFA.

An electronic questionnaire in the form of word documents was administered on a Chinese on-line survey system named "WenJuanXing" (https://www.wjx.cn/index.aspx) through their mobile phones for their convenience and for research efficiency, if they wished to use their phones. Data collected on the system were automatically stored as .sav files, a default format that could be analysed on SSPS. Data collection on the questionnaire for each student lasted approximately 10 to 20 minutes (Beatty et al., 2020; Dörnyei & Taguchi, 2010). Students performed the speaking test tasks on the headphones connected to computers with TPO software packages. Their responses to the speaking test tasks were recorded automatically by the software packages and were stored on the computers automatically as a single file. These files were named after the students' codes. The order of those recording files was randomised using a random list generated in the Microsoft Excel before they were given to the two raters. All the recording files were backed up in case of data loss (Weir et al., 2006).

4.6.2 Phase Two

Data collection in Phase Two was for investigating the research questions. It was broken into two parts: Data collection on students and on teachers.

4.6.2.1 Students

Although the quantitative data collection on the students in Phase Two followed almost the same procedure as in Phase One, differences existed in four aspects:

a) The students were different from those in Phase One.

b) In Phase One, the students only answered the questionnaire after they finished all the four speaking test tasks, whereas those engaged in Phase Two had to respond the questionnaire each time they finished one integrated speaking test task.

c) The questionnaire used in Phase Two was different from what was employed in Phase One. In Phase Two, the validated version of the questionnaire was used, whilst in Phase One the draft versions of the questionnaire was adopted.

d) The students' oral scores were formally rated by the two trained raters.

In Phase One, the students' speaking performance was only used for rater training and scoring piloting. By contrast, their speaking performance in Phase Two was formally rated for the subsequent data analysis. By means of analytic scoring before holistic scoring (see Section 3.4.4), two trained raters firstly scored independently the four segments of each student participant's prompts by referring to the rubrics. A score ranging from 0 to 4 points was given to the four segments. Then the four scores were aggregated to form a composite score for each participant's response on each task. The composite scores from the two raters for each response were then aggregated before they were divided to generate an average score which was used as the holistic score to measure the oral performance of the students statistically (Huang & Huang, 2013).

In the qualitative data collection regarding the semi-structured interviews, the students were interviewed in Chinese individually in a comfortable classroom after they were given the PISs and signed the CFs. The report language in the interviews was not

specified, as the interviewees were allowed to use whatever language (either English or Chinese) that they felt comfortable with. At the beginning of the interviews, a briefing was provided to the students such as the research objectives and relevant ethic issues. Enough time between questions was given for improving the interviewees' recollection of the past events, thereby increasing the validity of their responses. Note-taking, audio-recording and researcher diary were used to catch every detail of the interviewees' responses as triangulation. Closing comments with gratitude was offered to the participants, and a research report to those who expressed interest was promised (Creswell & Poth, 2018; Powney & Watts, 2018). Each individual interview lasted approximately 30 minutes and was audio-recorded for later transcription.

4.6.2.2 Teachers

Data on the teachers were collected via emails for their convenience. Before data collection, I had a face-to-face meeting with each of the teachers, during which the speaking test tasks and the teachers' role in the independent evaluation of task complexity were introduced, and their questions about this study were answered. All the teachers were suggested to do the speaking tasks before their independent evaluation for their acquaintance with the tasks. In the end, five valid responses were received, but one respondent did not answer the open questions on the self-rating questionnaire. All the teachers signed the PIS and the CF before participating in this study.

4.7 Data Analysis

Data analysis in Phase One was about instrument validation on the questionnaire. In Phase Two, data were analysed for answering the research questions.

4.7.1 Phase One: Questionnaire validation

Data analysis in Phase One was to validate the newly developed questionnaire, and to examine its reliability. "Validity is the extent to which an instrument measures what it is supposed to measure" (Kimberlin & Winterstein, 2008, p. 2278) and reliability indicates the stability of measures administered regardless of the time and the samples related to a particular study (Kimberlin & Winterstein, 2008). The overall reliability and the validity of a new questionnaire must be assessed with the use of factor analysis and its reliability can be further examined via internal consistency (Kline, 2016; Reinard,

2006; Qin, 2003). In line with this, EFA and CFA were run on the draft versions of the Metacognitive Strategy Inventory for its construct validity. To be specific, the EFA was conducted to determine the item sets that clustered in the questionnaire, and discover the common factor affecting a cluster of items on the questionnaire, whilst the CFA validated the factor structure extracted in the EFA (Zhang, 2014).

Three fundamental steps were involved in the EFA: (a) the examination of the feasibility of the dataset for EFA with reference to Bartlett's test of sphericity ($p< .05$) and the KMO ($p >.7$); (b) factor extraction; and (c) scale items loading on a particular factor. (b) and (c) were determined by factor extraction methods and factor rotation methods respectively (Brown, 2015; Byrne, 2016; Kline, 2016; Qin, 2003; Zhang, 2014). Maximum Likelihood (ML) estimation was adopted for factor extraction and Promax rotations were employed for factor rotation (Sun, 2016; Sun, Zhang & Gray, 2016; Vandergrift, et al., 2006). The factor loadings of each scale item on a factor were examined, and any items that had a factor loading below 0.4 or that loaded on more than one factors were removed from the draft questionnaire (Beavers et al., 2013; Brown, 2015; Byrne, 2016; Kline, 2016; Sun et al, 2016).

After the EFA, a model was extracted subject to further cross-validation on CFA (e.g., Qin, 2003; Teng & Zhang, 2016; Sun et al., 2016). The CFA was divided into two steps: Pre-CFA and the CFA. In the first step, model specification, model identification and assumption tests were conducted. Model specification was built upon the structure generated from the EFA (Byrne, 2016; Cohen, 2014; Kline, 2016; Oxford, 2017; Phakiti, 2003, 2018; Zhang, 2014, 2016, 2017; Zhang & Zhang, 2018). Model identification was conducted with reference to the guidelines proposed by Byrne (2016), and Kline (2016). These guidelines include (a) Scaling latent variables (the variance of the first indicator of factors was fixed to a value of 1. 0); (b) deciding on the number of parameters (the number of figures reflected by the input matrix should be not less than the number of freely estimated model parameters); and (c) deciding on the number of indicators of each latent variable (≥ 3). The examination of model fit was based on the fit indices including Goodness-of-fit (GFI), incremental fit index (IFI), Tucker-Lewis coefficient (TLI), comparative fit index (CFI), and the root-mean-square error approximation (RMSEA). The acceptable cut-off points for GFI, IFI, TLI and CFI were

greater than .9 and that for RMSEA was less than .8. After factor analyses, the reliability of the questionnaire was examined with reference to the Cronbach's alpha coefficient and the thumb-up criterion was over .8 (Brown, 2015; Byrne, 2016; Kline, 2016; Qin, 2018; Teng et al., 2018; Teng & Zhang, 2016). In the CFA, the estimation method of ML was employed as in the EFA.

Data analysis on the EFA was conducted in SPSS.24 (Qin, 2003; Reinard, 2006). Although two popular statistical packages can be used for the CFA (AMOS and LISREL), owing to the resources accessible, I validated the model generated from the EFA with AMOS.24 (Kline, 2016; Reinard, 2006; Qin, 2003).

4.7.2 Phase Two: Investigating research questions

Data analysis in Phase Two encompassed two stages: Quantitative analysis for Research Questions 1, 2, 4, 5, 6, 7, 8; and qualitative analysis for Research Question 3 and for a deep understanding of Chinese EFL learners' metacognitive strategy use.

4.7.2.1 Quantitative analysis

Quantitative analysis was further broken down into two steps run on two statistical software packages: SPSS. 24 and HLM.7. In Step One, basic information on the students' metacognitive strategy use across tasks, their oral scores, and task complexity was analysed to address Research Question 1, Research Question 2, and Research Question 4. Such information also served as the precondition of Step Two in which the relationships in the three variables with one another were investigated. The following subsections present the specific statistical procedure for each of the specific research question.

Descriptive analysis for Research Questions 1 & 4

As a preliminary analysis, descriptive analysis provides the basic information on the measures of the central tendency including mean, median and mode, and the measures of variability such as standard deviation, skewness and kurtosis. The mean (the average) is the most used method by researchers to describe the central tendency of a specific dataset, whereas the skewness and the kurtosis are the key parameters in normality inspection (Muijs, 2011; Pallant, 2016; Peng, 2009). Due to this, descriptive analysis

has been adopted by some researchers in analysing the participants' strategy use and their speaking test scores (e.g., Youn & Bi, 2019; Sun, 2016). In line with them, I referred to the means of the metacognitive strategies reported by the Chinese EFL learners, and of their oral scores on the four integrated speaking test tasks generated in the descriptive analysis to address Research Question 1 and Research Question 4. In the meanwhile, the standard deviation, the values of the skewness and kurtosis were inspected for normal distribution of the datasets in the assumption testing.

Correlation analysis and one-way repeated measures ANOVA for Research Questions 2 & 6

Pearson product-moment correlation is used to measure correlations between two variables when the datasets on the variables are normally distributed (Mujils, 2011). This statistical method was employed to examine the relationships between task difficulty and mental efforts, the two items on the self-rating scales, after normality inspection. The strong correlation ($p \leq .05$) between the two items suggested that task difficulty could indicate task complexity statistically (Révész et al., 2016; Robinson, 2005). Given that the data on the students were collected from repeated observations nested within them, the variability in the students' perceptions of task complexity of the integrated speaking test tasks was investigated on a one-way repeated measures ANOVA to address Research Question 2 (Muijs, 2011; Pallant, 2016). In the procedure, task complexity was treated as the independent variable represented by the four integrated speaking test tasks and task difficulty was the dependent variable. The means of task difficulty were compared to examine the variance within the four tasks (Frey, 2018; Révész, 2014; Sasayama, 2015, 2016). Indices, which indicated significant variance in task complexity across tasks, were inspected. They included the *p*-value for the F-ratio ($p \leq .05$), and the η^2 which suggested the effect size: If η^2 is $\leq .01$, it suggests a small effect size; a value ranging from .01 to .06 indicates a moderate effect size, and if η^2 is $\geq$ .14, it indicates a larger effect size (Pallant, 2016).

By the same token, a series of one-way repeated measures ANOVA were run on the dataset of task difficulty and the students' oral scores to test the correlations between the two variables for Research Question 6. In the procedure, task difficulty, the indicator of task complexity as explained above, served as the independent variable or the

treatment conditions and the students' oral scores were treated as the dependent variables (Allen, 2017; Frey, 2018; Pallant, 2016; Rahimi, 2016).

Multiple regression analysis for Research Question 5

To research the relationships between the students' metacognitive strategy use and their speaking performance for Research Question 5, multiple linear regression was deployed (Allen, 2017; Frey, 2018; Pallant, 2016). The statistical procedure was used to assess how the four individual metacognitive strategies clustered to explain the students' speaking performance while examining the associations between individual metacognitive strategies and test performance. The four subcomponents of the metacognitive strategies were entered into a model simultaneously as the predictor variables and the students' oral scores were entered into the model as the outcome variable. Correlation coefficients (r) within the four individual metacognitive strategies were examined first for the appropriateness of the statistical procedure, and for inspecting multicollinearity: When r is $\leq .8$, the employment of the procedure is suitable. Index regarding model fit was the adjusted R^2, and the rule of thumb for the index is presented as the following:

< 0.1: poor fit
0.11 - 0.3: modest fit
0.31 - 0.5: moderate fit
> 0.5: strong fit

In addition, as the four strategies were measured on the same units on the questionnaire, the unstandardized coefficients (β) were examined to investigate the impact of each individual metacognitive strategy on the students' speaking performance. The cut-off p-value for β parameters is < 0.01, indicating substantive effects of a specific metacognitive strategy on the students' speaking performance (Allen, 2017; Frey, 2018; Pallant, 2016; Xie, 2013).

One –way repeated measures MANOVA for Research Question 7

Since the construct of metacognitive strategy had four individual components and task difficulty had four task conditions, one way repeated measures MANOVA was used to

examine the correlations between the two variables for Research Question 7 (Allen, 2017; Frey, 2018; Pallant, 2016). A new variable that combined the four individual metacognitive strategies linearly was created to investigate the within-subject variance in the students' reported use of the clustering metacognitive strategies across tasks. For identifying variance, values of F ($p < .05$) and η^2 were examined, and the rule of thumb-up for these indices were the same as that of ANOVA presented above. The exact location of the variance in the four individual metacognitive components was further detected via ANOVA (Frey, 2018; Muijs, 2011; Pallant, 2016).

Hierarchical modelliing for Research Question 8

Although hierarchical data structure, particularly the structure of repeated-measures data (e.g., observations on test-takers shown by their speaking performance nested within the test-takers) as the case with this study is very common in L2 assessment research, the most widely used statistical techniques to investigate such data structure are single-level techniques: ANOVA, multiple regression analysis (MR), G-theory and multi-faceted Rasch models (MFRM). In consequence, incorrect results may be produced such as biased standard errors and confidence intervals (Barkaoui, 2013; Raudenbush & Bryk, 2002). To avoid such statistical errors, a two-level Hierarchical Linear Model (HLM) was established in investigating the relationships among the three research variables for Research Question 8 (Barkaoui, 2013). In the model, Level-1(task level) involved task conditions reflecting task complexity. At this level, the students' oral scores were the outcome variable and the four task conditions established by the four integrated speaking test tasks were the predictor variables. At Level-2 (student level), the predictor variables were the metacognitive strategies reported by the Chinese EFL learners (e.g., Barkaoui, 2010, 2013; Wen, 2009; Weng, & Qiu 2014; Raudenbush & Bryk, 2002; Snijders, 2012). One thing must be pointed out that HLM in this study refers to both the statistical procedure and the statistical package (Barkaoui, 2010, 2013; Garson, 2013; Wen, 2009; Weng, & Qiu 2014; Raudenbush & Bryk, 2002).

Data analysis of the model was performed in two steps. Step One focused on the effects of the interaction between the four individual metacognitive strategies and task complexity on speaking performance. In Step Two, the impacts of the interactions of the metacognitive strategies in both the separate and the interactive manners with task

complexity on speaking performance were investigated simultaneously. To model the interactions within the metacognitive strategies, the four individual strategies were multiplied with one another for new interaction variables (Muijs, 2011; Pallant, 2016; Young, 2017). For instance, to model the interaction between planning and problem solving, data on the two strategies were multiplied with each other. In accordance with this rule, a total of 11 new variables indicating the interactive metacognitive strategies were created as shown in Table 4.7.

Table 4.7 *New Variables at Level-2 in Step Two*

Name	Interaction Terms
V1	Planning × Problem solving
V2	Planning × Monitoring
V3	Planning × Evaluating
V4	Problem solving × Monitoring
V5	Problem solving × Evaluating
V6	Monitoring × Evaluating
V7	Planning × Problem solving × Monitoring
V8	Planning × Problem solving × Evaluating
V9	Problem solving× Monitoring × Evaluating
V10	Planning × Monitoring × Evaluating
V11	Planning ×Problem solving ×Monitoring × Evaluating

Note. V=Variables.

Due to different research foci, the predictor variables at Level-2 varied despite the same variables at Level-1 in the two research steps. In Step One, the four independent metacognitive strategies were Level-2 predictor variables. In Step Two, however, Level-2 predictor variables involved not only those in Step One, but the interactive metacognitive strategies reflected by the 11 newly generated variable shown in Table 4.7.

Following variable identification was data preparation which included assumption testing, feasibility testing, model building and evaluating, and model checking. In data preparation, data were transformed from wide form into long form for MDM files, the

default format of the statistical package for the HLM. After data preparation, assumption tests and feasibility test to examine whether HLM was appropriate for the current dataset were conducted. In model building, the dataset of metacognitive strategies was entered the model after it was grand-mean centred and that of task difficulty was group-mean centred. Two models were established in both of the two research steps: One was the null model without any predictor variables to assess whether a HLM model was needed with reference to the Intra-class Coefficient or ICC: A high ICC indicates the necessity of data analysis on HLM. In addition, the null model served as the benchmark value of the deviance for model comparison in model building. The other model was the full model with all the variables being entered into the model simultaneously to study how metacognitive strategies and task complexity interacted, and how such interaction exerted impacts on speaking performance. Put it differently, this entry approach examined the cross-level interactions. Full maximum likelihood estimation (FML) was adopted in model evaluation. Two main indices were inspected for model fit: Deviance statistics (the decrease in the value indicates better model fit), and significance tests which included the t-tests for the fixed effects ($p < .05$) of the testing parameters, and the Chi-square tests to examine those parameters' random effects ($p < .05$). The reliability of Level-1 random coefficients was also inspected because it represented the ratio of the differences in the individual characteristics including their use of the metacognitive strategies that attributed to the true individual variability in the test scores (Barkaoui, 2010, 2013; Garson, 2013; Raudenbush & Bryk, 2002).

In the diagnostic process which also served as part of the assumption testing, multicollinearity at Level-2 was diagnosed with visual inspection of nonlinearity by means of scatter plots of ebintrcp (Empirical Bayes intercept estimate) or EBINTRCPT1 from Level-2 residual file against Level-2 predictor: If there are no oval-shaped scatter plots identified, nonlinearity does not exist in Level-2 residuals. The normality of the residuals of Level-1 and Level-2 were examined with reference to scatter plots. If the scatter plot of the chipct (the expected values on the chi-square distribution) against midst (Mahalanobis distance) from Level -2 residual file is a 45-degree line, the assumption of the normality at Level 2 is met. Similarly, heteroscedasticity or outliers was examined via the visual inspection of the scatter plot of EB (Empirical Bayes) residuals against the predicted values from Level-2 residual file. In addition, the robust

standard errors were compared with the ordinary standard errors for any possible model misspecification: A substantical difference between the two suggests model misspecification (Barkaoui, 2010, 2013; Garson, 2013; Wen, 2009; Weng, & Qiu 2014; Raudenbush & Bryk, 2002).

In the investigation into the assumed relationships among metacognitive strategies, task complexity and speaking performance, a simultaneously examination of the effects of task complexity on test performance and the effects of the Chinese EFL learners' metacognitive strategy use on their speaking performance was conducted in the HLM. In this sense, the statistical procedure not only addressed Research Question 8, but to cross-validate Research Question 5 and Research Question 6 as well (Allen, 2017).

4.7.2.2 Qualitative analysis

Qualitative data analysis was divided into two stages. The first stage aimed at analysing the data generated from the semi-structured interviews and the second stage related to those collected on the open-ended questionnaires.

The semi-structured interviews

Coding Methodology

Data collected from students were subjected to structure coding followed by content analysis in a deductive approach for their perceptions of metacognitive strategy use and task difficulty. The employment of the coding methodology was due to the use of the semi-structured interviews and the open-ended questionnaires, and the clear presentation of the research questions (see Section 1.6.3). Data transcription and analysis for each interview followed the same guideline.

Structure coding "applies a content-based or conceptual phrase representing a topic of inquiry to a segment of data that relates to a specific research question used to frame the interview" (Saldana, 2016, p. 84). Saldana (2016) suggests that structure coding is particularly appropriate for semi-structured interviews and open-ended questionnaires. Structure coding is typically followed by content analysis. Content analysis, as Drisko and Maschi (2015) propose, is appropriate for datasets generated from newly developed

interviews, and it is particularly useful when the research question is clearly defined, and the categories are established by existing studies as the case with this study

Coding Scheme

In line with the research questions, the semi-structured interview scheme (see Table 4.1), the working definitions of the metacognitive strategies (see Table 3.1), and task complexity factors (see Table 3.2), a coding scheme was developed. In the scheme, metacognitive strategies and task complexity were used as the two broad categories, the four constructs of the metacognitive strategies and the five task complexity factors (the overall measurement of task complexity was the fifth factor in addition to the four factors) served as the sub-constructs of the coding themes. The definitions of the subcomponents of the metacognitive strategies were used as the codes for coding the students' metacognitive strategy use. On the other hand, the presumed students' attitudes (e.g., in support of longer versus shorter planning time) toward the task complexity factors based on Robinson's (2015) Triadic Componential Framework were employed as the codes for task complexity. In total, there were 17 codes on metacognitive strategies and 10 codes on task complexity generated as shown in Table 4.8.

Table 4.8 *Coding Scheme*

Categories	**Subcategories**	**Codes**	**Definitions**
Metacognitive strategies	Planning	Setting goals	Identify the purpose of the task
		Directed attention	Decide in advance to focus on the given tasks and ignore distractions
		Activate background	Think about and use what you already know to help you do the task
		Prediction	Anticipate information to prepare and give direction for the task
		Organizational planning	Plan the task and content sequence
		Self-management	Arrange for conditions that help you learn
	Problem-solving	Inference	Make guesses based on previous knowledge
		Substitute	Use a synonym or descriptive phrase for unknown words
	Monitoring	Selective attention	Focus on key words, phrases, and ideas
		Deduction/induction	Consciously apply learned or self-developed rules
		Personalize/personal experience	Relate information to personal experiences
		Take notes	Write down important words and concepts
		Ask if it makes sense	Check understanding and production to keep track of progress and identify problems
		Self-talk	Talk to yourself to reduce anxiety
	Evaluating	Verify predictions and guesses	Check whether your predictions or guesses are correct
		Evaluating performance	Judge how well you do in the task
		Check goals	Decide whether goal was met
Task Complexity	Planning	Positive on shorter time	Shorter planning time better for task execution
		Positive on longer time	Longer planning time better for task execution
	Prior knowledge	Attitudes to campus life	Campus life situations as prior knowledge provider
		Attitudes to academic lectures	Academic lectures as prior knowledge provider
Task Complexity	Steps involved	Reading as prior knowledge	Reading provides background information

Categories	Subcategories	Codes	Definitions
		Reading as extra cognitive load	Reading imposes extra information load
	Task types	Positive attitudes	Task types influence positively on task performance
		Negative attitudes	Task types affect negatively on task performance
	Overall judgment	easy	Overall, the task is easy
		difficul	Overall, the task is difficult

Coding Process

Given the large dataset, analysis of the students' responses to the interview guide was conducted on the software package of NVivo 12 (Windows version) (Bazeley, 2012). NVivo supported the direct import of the documents from a word processing package, which simplified coding process on the computers and improved research efficiency (Bazeley, 2012; Saldana, 2016). Relevant literature also justified the use of NVivo in processing semi-structured interviews (Bazeley, 2012; Saldana, 2016; Sun, 2016).

Before coding was data preparation which started with the transcription of the interviews verbatim to word documents. The transcripts was then imported into NVivo on which nodes representing the eight categories and the sub-nodes reflecting the two sets of codes were created. As the data were clearly structured, I first automatically coded all the interviewees' responses to a particular question using the question as the heading. The automatic coding was in line with Bazeley's (2012) suggestion that researchers should code responses to structured questions with automatic coding on NVivo. As a result, the interviewee's responses were organized orderly under the questions in the interview guide, which simplified the ensuing detailed coding, improving coding efficiency. In detailed coding, I examined the transcripts line by line and word by word with reference to the coding scheme shown in Table 4.9 for analytic memos which were used later in combination with codes for themes.

The methods of vertical and horizontal analyses were employed in coding. Vertical analysis was utilised to investigate the individual interviewee's responses separately with special attention to extreme cases. After the vertical analysis of individual cases, horizontal analysis started for the commonalities and differences in the individual interviewee's reported use of metacognitive strategies and their independent judgments of task complexity in a cross-case manner (Richards, 2015). Common themes, recurring patterns and uniqueness of any particular case were noted to see if they were within the two sets of codes in the coding scheme. Regards those that were beyond the scheme, I coded them as new ones in accordance with relevant literature on metacognitive strategies and task complexity. For instance, during the actual coding process, I found when asked whether they set goals before the speaking tasks, some students reported that they did not set any goals during planning time, so I coded their response as "have no particular plans". This code was not originally included in the coding scheme. Final

data coding involves reexamination of the data to reassess the coding for refinement (Cresswell & Poth, 2018; Paltridge & Phakiti, 2010). The coding process is demonstrated by Figure 4.4.

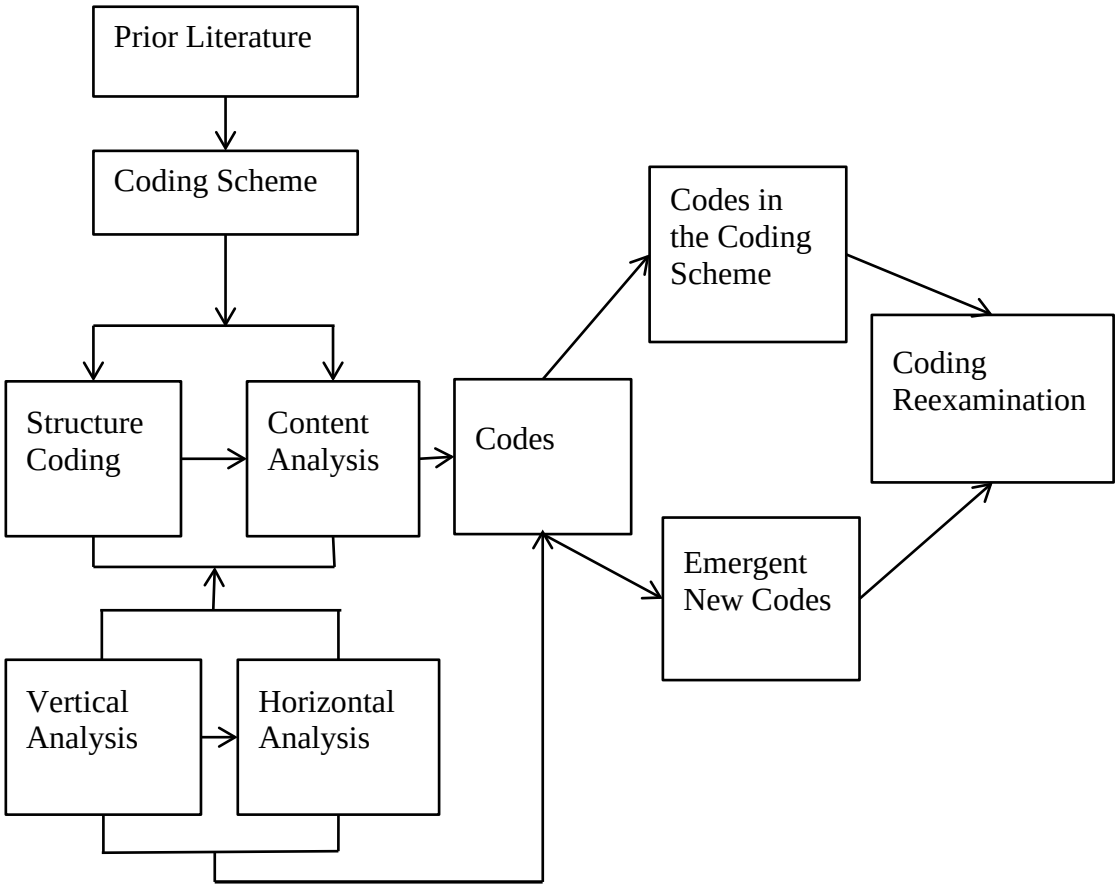

Figure 4.4 *Coding Process*

An example of data presentation and data coding

Direct quotes from the interviewees were used to present the qualitative findings. Since the vast majority of the quotes were Chinese with a few involving English expressions, they were translated into English with the consultation of the two experts from SISU for the credibility of the presentation. At the end of each quote, data source was shown in the brackets. For example, the interviewee who participated in the second interview on 24th of June in 2018 was presented as Interviewee 2/24/06/2018. An example of the quotes and the coding is presented as the following:

> When I got some words that I didn't understand, I made a guess …guess why it was used according to the context…, and this method worked on me…. I often

did this when I got problems in taking the speaking tasks. (Interviewee 2/24/06/2018)

In this unit of data, Interviewee 2 reflected what she did when she got problems in performing tasks in response to the interview prompt which asked her what she did to solve problems. Via structure coding, the segment was coded as "solving problems in performing tasks" and put it in the node called "problem solving" on NVivo. In the subsequent content analysis, this unit of data was put into a sub-node named "make a guess on the basis of contexts or prior knowledge", which was later to be further themed as "inference" with similar responses from other interviewees in light of the coding scheme on the metacognitive strategies.

Then open-ended self-rating scales

Data analysis of the teachers' measurement of task complexity produced from the open-ended questionnaire followed the similar procedure on the students as presented above. However, given the comparatively small datasets of the teacher participants (N = 5), I analysed the data manually and independently in line with Révész et al. (2016). Guided by the deductive approach with content analysis, I adopted the descriptive coding method in which simple words and phrases were used to summarise each teacher's answers before annotating them in a particular category with reference to the coding scheme on task complexity via horizontal and vertical analysis. Finally, a frequency count of all the annotations for the four tasks was obtained after the annotations were added up and fell into a specific category (Révész et al., 2016).

4.8 Minimising Risks to Validity and Credibility

Due to the research design, there might be risks to the validity and credibility of the research findings. To minimise the risks, the following methods were used.

4.8.1 Data alignment

In the overall mix-methods design, poor alignment in mixing the data from the Metacognitive Strategy Inventory and those from the Semi-structured Interview Guide may impose risks to the validity of the research results. To minimise such risks, the prompts on metacognitive strategies utilised in the interview guide were highly similar

to the items of the questionnaire (see Section 4.5). The similarity attributed to the high level of agreement between the two instruments, and hence for the general validity of the present study. To further ensure the alignment of the quantitative data and the qualitative data, the collection of the two sources of data was separated by only a short period of time. The adoption of simple internal structure in both the questionnaire and the interview guide in concert with the presentation of the items and the prompts in a concrete and specific way also enhanced data alignment (Harris & Brown, 2010).

4.8.2 Counterbalancing carryover effect and order effect

The carryover effect and the order effect in the one-way repeated measures design might impose risks to the validity of this study. Carryover effect refers to the effects that a previous testing condition may have on participants such as fatigue. To minimise the effect, the students were offered 20- minute of interval between each of the four integrated speaking test tasks, during which they were provided with some refreshments. The order effect was generated from the order of the four integrated speaking tasks. To minimise the risk, counterbalancing was utilised via a Latin square design (Corriro, 2017; Gilabert, 2007a, 2007b; Verma, 2015; Weir et al., 2006), in which the students were randomly divided into four groups in light of the sequence of the integrated speaking tasks. Each group had almost equal number of the students who simultaneously performed the four tasks in different sequence. For example, Group 1 stood for the task order of Task 3, Task 4, Task 5 and Task 6; Group 2 represented the sequence of Task 4, Task 5, Task 6, and Task 3, and so on so forth. If a student was in Group 1, he or she performed the four tasks in the order of Task 3, Task 4, Task 5 and Task 6. Likewise, if a student participant was in Group 2, he or she finished the four tasks in the order of Task 4, Task 5, Task 6, and Task 3 (Verma, 2015; Vogt, 2011). Table 4.9 displays the task order of the four speaking tasks for the four student groups.

Table 4.9 *Tasks in Sequence Based on a Latin Square Design*

Group	Sequence of Tasks
Group 1	Task 3/ Task 4/Task 5/ Task 6
Group 2	Task 6/Task 3/ Task 4/Task 5
Group 3	Task 5/ Task 6/Task 3/Task 4
Group 4	Task 4/ Task 5/Task 6/ Task 3

4.8.3 Triangulation, member checking, and auditing

Risks to the accuracy of the qualitative findings were reduced through triangulation, member checking, and auditing (Creswell & Guetterman, 2019; Creswell & Poth, 2018). The combination between quantitative and qualitative data generated from various sources (the students and the teachers) by multiple means (observational field note, recorded responses and research diary) attributed to triangulation (Révész et al., 2016). The engagement of several interviewees in examining the raw data, narrative account, and the interviewers' comments to see if their real ideas were accurately reflected contributed to member checking. The participation of an external auditor in a thorough review of this research led to external auditing (Brinkmann & Kvale, 2015; Creswell & Poth, 2018).

4.8.4 Intra-rater reliability and inter-rater reliability

Risks from raters related to intra-rater reliability was mitigated via rater training (Huang et al., 2013; Huang et al., 2018; Sun, 2016), and inter-rater reliability were achieved with reference to the Cronbach's alpha coefficient as stated previously (Ergai et al., 2016; Hallgren, 2012; Sun, 2016; Upton & Cook, 2014).

4.9 Ethical Considerations

As this study involved the students and the teachers, ethical issues were addressed properly. It was approved by the University of Auckland Human Participants Ethics Committee (Reference Number 020972).

Before the study was conducted, I contacted relevant departments in the research sites for official permission. An informed CF was provided, in which the purpose of the

research and the information indicating that participants were voluntary, and the research would never place them at undue risks were clearly stated. All the participants were provided with the PISs and the CFs for their signatures, and they were entitled to ask me to destroy the data collected unconditionally. Furthermore, they had the right to refuse to answer any specific questions, and to ask the stop of the recording and field note during the interviews. The participants engaged in the qualitative data collection were given a small gift worth around $20NZD as a token of gratitude. A thank-you letter was provided for all the participants.

Deans of the relevant faculties in the research sites also signed the CFs and the PIS, stating that the participants' participation or withdraw would not cause any consequence to them or to anyone at any level of the faculty; and participation or withdraw would not affect the participants' relationship to the faculties and their academic credits.

The participants' anonymity and confidentiality were protected. When answering the questionnaire, the students were asked to write a code name only recognised by them. Similarly, the teachers used a unique code name to answer the open-ended questionnaire. As such, no individual university, student or English teacher was identifiable, and those engaged in the study such as the raters and the research assistants were invited to sign a Confidentiality Agreement to protect the participants' confidentiality.

With regard to data management, the hard copy data was securely stored in a locked cabinet at the University of Auckland, and the electronic data was stored confidentially in my computer before they are destroyed six years after the ethics approval. The data collected might be used for academic purposes in the form of publications or conference presentations with the participants' permission (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Creswell & Poth, 2018).

Chapter Five
Quantitative Results

5.1 Chapter Overview

This chapter reports the quantitative results in the two phases. In Phase One, results of the exploratory factor analysis (EFA) and the confirmatory factor analysis (CFA) are reported for the validation of the Metacognitive Strategy Inventory (MSI). In Phase Two, results of various statistical tests are presented to address the research questions.

5.2 Phase One: Instrument Validation

As stated above, the instrument validation had to do with the EFA and the CFA. Results of the two factor analyses were presented sequentially in the following subsections.

5.2.1 Exploratory Factor Analysis (EFA)

Results of the EFA were reported in two steps: Assumption tests and several rounds of factor analyses. Assumption tests involved handling missing data, normality, outliers, linearity, multicollinearity, and evaluating factorability via Bartlett's Test of Sphericity and Kaiser-Meyer-Olkin test of sampling adequacy. On the other hand, the five rounds of factor analysis extracted four factors with 23 items for further examination in the CFA.

5.2.1.1 Assumption tests

Descriptive analysis revealed that there were no missing data for the dataset of the 40 items of the original MSI. Values of the skewness of these items were between -. 018 and. 427, and figures for kurtosis ranged from - .902 to .273, with all the values falling within the acceptable bounds for univariate normality (Beavers et al., 2013; Kline, 2016) as shown in Table 5. 1.

Table 5.1 *Descriptive Analysis of the 40 Items*

Item	Means	Skewness	Kurtosis	S. D
Q1	4.21	-.180	-.903	.081
Q2	3.94	-.077	-.771	.082
Q3	3.37	.299	-.246	.077
Q4	3.57	.022	-.548	.071
Q5	3.69	.099	-.855	.074
Q6	3.64	.186	-.744	.073
Q7	3.67	.099	-.629	.075
Q8	3.11	.373	-.149	.070
Q9	3.11	.427	-.023	.072
Q10	4.43	-.408	-.723	.082
Q11	3.63	.117	-.579	.077
Q12	3.61	.034	-.494	.074
Q13	3.78	-.077	-.584	.080
Q14	4.07	-.178	-.156	.068
Q15	4.11	-.202	-.298	.067
Q16	3.02	.248	-.682	.086
Q17	3.96	-.043	-.393	.075
Q18	3.15	.227	-.634	.081
Q19	3.96	-.071	-.373	.065
Q20	3.89	.036	-.438	.067
Q21	3.83	.019	-.347	.070
Q22	4.39	-.300	-.663	.071
Q23	3.31	-.108	-.384	.068
Q24	3.59	.139	-.231	.072
Q25	3.75	.175	-.432	.071
Q26	3.54	-.050	-.344	.077
Q27	3.06	.206	-.509	.079
Q28	3.48	.080	-.429	.080
Q29	3.30	.266	-.150	.070
Q30	3.61	.256	-.508	.075
Q31	3.32	.194	-.550	.074
Q32	3.76	.196	-.432	.071
Q33	3.12	.562	-.015	.062
Q34	3.60	.227	-.506	.072
Q35	3.63	.022	-.495	.077
Q36	3.65	.224	-.217	.067
Q37	3.46	.271	-.211	.070
Q38	3.60	.124	-.450	.076
Q39	3.49	.220	-.435	.075
Q40	3.48	.286	-.155	.069

Note. Q = Question; S. D = standard deviation.

Visual inspection of the histograms with normality curve, Q-Q plots, and box-and-whisker plots provided additional confirming evidence of the univariate normality of the dataset. Based on the univariate normality test, Z scores of the items were calculated, and all met the cut-off value of +/-3 for identifying outliers. This result and the examination of the box plots suggested that no univariate outliers existed. However,

multivariate outliers (N = 30) were discovered. After the removal of these outliers, the final sample size for the EFA was 224 participants for a 40 item scale, meeting the thumbs-up rule: The subject-to-variables ratio should be 5:1. The subsequent regression analysis on multicollinearity displayed that values of Tolerance of the 40 items were all above the cut-off point of .2, and the numbers of their variance inflation factor (VIF) were all less than 5, the cut-off boundary. The results indicated that multicollinearity did not exist. Given the rather large number of items in the MSI, linearity was examined between the item with the strong negative skewness (Item 10) and the item with the strong positive skewness (Item 33) via a scatterplot (see Table 5.1). The examination disclosed the multivariate normality of the dataset (Beavers et al., 2013; Field, 2018; Kline, 2016; Salkind, 2010; Scott, 2014; Tabachnick & Fidell, 2013).

To evaluate the factorability of the data, the Bartlett's Test of Sphericity and Kaiser-Meyer-Olkin (KMO) test were inspected via initial factor analysis. Results showed that the strength of the relationships between variables was statistically significant: $\chi 2$ (df = 780) = 4740.273, and $p < .001$. The significance evidenced that the 40 items on the MSI were statistically sufficient for an EFA procedure. In addition, the value of KMO was.924, suggesting that the sample size was adequate for the EFA (Beavers et al., 2013; Brown, 2015; Byrne, 2016; Field, 2018; Kline, 2016).

5.2.1.2 The EFA

In the initial round of the EFA, the extraction method of Maximum Likelihood (ML) was administered on the 40 items. With reference to the eigenvalues, the scree plots and the percentage of variance extracted, eight factors were extracted, which explained 62.159 % of the total variance. However, on one of the extracted factors, there were 39 items with their factor loadings all above the cut-off value of .4. After the rotation of the extracted factors via the method of Promax, numbers from the Pattern Matrix showed that items with factor loadings above .4 scattered among eight factors. Despite this, none of the factor had at least three items (the cut-off criterion), indicating the failure of extracting the underlying common factors from the 40 items. Considering the parsimony and the meaningfulness of the eight-factor solution based on the working taxonomies of the metacognitive strategies (see Table 3.1), an alternative approach was employed to extract factors: The number of factors and their name were determined prior to the

extraction (Brown, 2015; Byrne, 2016; Kline, 2016; Qin, 2003). Accordingly, four factors were generated: Planning, problem-solving, monitoring and evaluating.

After the first round of the EFA on the four-factor solution via ML extraction and Promax rotation, the four factors only explained 49. 96 of the total variances, and indices of the model fit (Goodness-of-fit) of this solution [$\chi 2$ (df = 626) = 1105.671, $p \leq .001$] did not demonstrate significant improvement compared with the eight-factor solution [$\chi 2$ (df = 488) = 700. 17, $p \leq .001$]. In the meanwhile, values in the Pattern Matrix showed that factor loadings of six items were less than .4 on any of the four factors. After the exclusion of these items in the second round of the EFA, dramatic improvement was seen in the model fit: $\chi 2$ (df = 321) = 590, $p \leq .001$, and the total variance explained by the four factors increased to 54.94%. Four undesired items were identified in this around of the EFA with their factor loading less than .4. The third round of the EFA was conducted following the removal of the undesired items, and the retained model was further improved: $\chi 2$ (df = 296) = 555.51, $p \leq .001$. The total variance explained by the four factors reached 55.75%. Only one item then had a factor loading of less than .4 which was deleted before another round of the EFA. Following the same procedure, a total of 29 items were extracted out of the original 40 items on the four factors. The four factors contributed to approximately 55.76% of the total variance: Factor One (planning) accounted for 38.33% of the variance, the largest contribution in the four factors; and the proportions of the variance contributed by Factor Two (monitoring), Factor Three (problem-solving), and Factor Four (evaluating) were 4.58 %, 5.97 % and 6.88 % respectively. Unexpectedly, one item fell on the wrong factor: Q22 (Question 22) was predicted to fall on the factor of “monitoring”, but it fell on the factor of “problem solving”. In this stage, the item was included in the extracted structure until further reliability testing was conducted.

Reliability analysis following the EFA was composed of subscale reliability analysis and the examination of full-scale reliability with reference to the Conbach coefficient, and the rule of thumb-up is above .7 in the research area of education (Field, 2018; Kline, 2016; Qin, 2003). In this study, the Cronbach coefficients for both the subscale and full-scale reliability were above .8, indicating good consistency within each factor and within the MSI. As Q22 did not fall on its expected factor, special attention was paid to this item in the subscale reliability test. Result showed that the deletion of the

item increased the overall reliability of the MSI from .941 to .942. It evidenced the exclusion of Q22, and the outcome of the follow-up EFA without the item also supported the removal, as the value of the total variance explained by the four factor-solution experienced moderate growth to 56.69 %. The proportions of the variance explained by the four factors after the exclusion were 39. 23% (planning), 6.66% (monitoring), 6.14 % (problem solving) and 6.14 % (evaluating). Model fit indexes were also increased: $\chi 2$ (df = 272) = 526. 27, $p \leq .001$. The result indicated a better structure without Q22. Therefore, after five rounds of extractions and rotations, a structure composed of 28 items underpinned by four factors was established. The factor loadings of the items and the internal and the overall reliability are reported in Table 5.2.

Table 5.2 *The Metacognitive Strategy Inventory after the EFA*

		Factor Loadings				
Factors	**Items**	**P**	**PS**	**M**	**E**	**α**
P	Q1	.521				
	Q2	.513				
	Q3	.670				
	Q4	.809				.886
	Q5	.734				
	Q6	.626				
	Q7	.523				
	Q8	.683				
	Q9	.679				
PS	Q14		.643			
	Q15		.645			
	Q17		.740			.845
	Q19		.701			
	Q20		.719			
M	Q23			.563		
	Q24			.430		
	Q26			.473		
	Q27			.610		.871
	Q28			.474		
	Q29			.787		
	Q30			.505		
	Q31			.625		
	Q33			.628		
E	Q35				.621	
	Q36				.595	
	Q37				.677	.859
	Q38				.880	
	Q39				.654	
Overall reliability		.941				

Note. P = planning; PS = problem- solving; M = monitoring; E = evaluating; α = Cronbach's alpha.

In addition, the output of the Component Correlation Matrix shown in Table 5.3 revealed moderate inter factor correlations (≥ .3 but ≤ .8), suggesting that the approach of Promax rotation run in this dataset was appropriate (Field, 2018; Qin, 2003).

Table 5.3 *Component Correlation Matrix for the Four Factors*

Factor	Planning	Problem- Solving	Monitoring	Evaluating
Planning	1.000	.689	.556	.585
Problem- Solving	.689	1.000	.560	.660

5.2.2 Confirmatory Factor Analysis (CFA)

Results of the CFA were reported in two sections: Pre-CFA and the CFA. Pre-CFA focused on model specification, model identification and assumption tests, while the CFA presented the examination of the offending estimate test followed by the assessment of model fit and model modifications. Validity and reliability evaluation were performed on the final model in the CFA.

5.2.2.1 Pre-CFA

Model specification and model identification.

After model specification and identification, a zero-order model (Model A) composed of four correlated factors was established. In the model, variance of the first indicator of each of the four factors was fixed to 1 by default on AMOS. Based on the formula of 1 / 2 [P (P + 1)] where P refers to the number of the items of the MSI, after the EFA (P = 28), the number of the parameters in the matrix was 406, greater than that of freely estimated model parameters (62). Moreover, each of the four factors had more than three indicators (nine indicators for planning and monitoring, and five indicators for problem solving and evaluating). Each indicator was constrained to only one factor with error terms associated with each indicator variable uncorrelated (Brown, 2015, Byrne, 2016; Kline, 2016; Reinard, 2006, Rong, 2011). Figure 5.1 displays the structure of Model A.

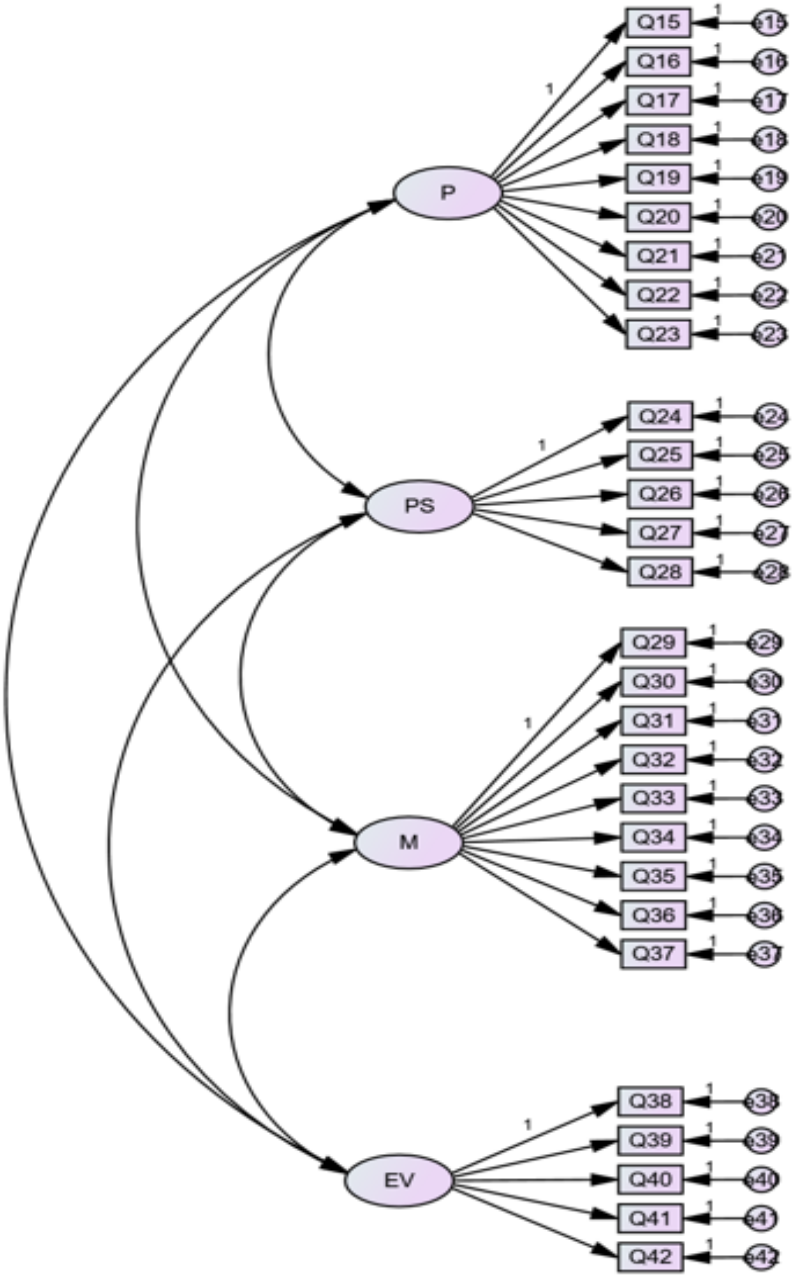

Figure 5.1 *Model A*

Note. P = planning; PS = problem solving; M = monitoring; EV = evaluation.

Here, it should be noted that Q15 was the default name of the variable for Item 1 in the MSI generated by AMOS in Model A. Similarly, Q16 was the default name of Item 2, and so on so forth. Table 5.4 provides the default names of each item of the MSI in Model A and their respective contents.

Table 5.4 *Default Names of the MSI Items in Model A and Their Contents*

IMs of the MSI	**DNs in Model A**	**Contents**
Q1	Q15	答题前，我注意了答题的要求。
Q2	Q16	我清楚题目要求我做什么。
Q3	Q17	我明白需要规划答题过程。
Q4	Q18	我想过需要做什么才能完成好答题。
Q5	Q19	我想过需要怎么做才能完成好答题。
Q6	Q20	我确信已清楚任务目标。
Q7	Q21	我明白完成任务所需的主要步骤。
Q8	Q22	我事先组织好了想说的内容的结构。
Q9	Q23	我对想说内容的词语与表达事先做了准备。
Q10	Q24	我会借助已有知识(如上单词的下文，构词及话题)来猜测陌生单词或词组的意思。
Q11	Q25	我根据上下文猜测话题。
Q12	Q26	我借助已有知识来完成话题任务。
Q13	Q27	(口语表达)想不起来某个英语单词时，我用相通常思的其他词或词组。
Q14	Q28	我使用近义词之类其他方法来表达意思。
Q15	Q29	我知道答题时何时该加快速度。
Q16	Q30	我知道答题中何时要更加仔细。
Q17	Q31	答题过程中我知道自己用掉多少时间。
Q18	Q32	口试过程中，我知道自己哪些地方说得比较地道。
Q19	Q33	口试过程中，我能意识到自己犯的语法错误。
Q20	Q34	我能把题中的信息与己有知识联系起来。
Q21	Q35	答题过程中我会记录重要词语与概念。
Q22	Q36	考试时我不断核查做好的题目和答题进度。
Q23	Q37	如果进展不顺利，我知道该怎么对付。
Q24	Q38	完成口语任务后我在脑子里给自己的表现打了个分数。
Q25	Q39	任务结束后我检查自己是否达到目标。
Q26	Q40	考试后我会检查自己的错误。
Q27	Q41	我会评价自己在答题表现方面的满意度。
Q28	Q42	我会评价原定计划实施的有效性。

Note. IMs of the MSI = items of the Metacognitive Strategy Inventory; DNs = default names; Q = questions.

Assumption tests

Model specification and identification were followed by assumption tests which involved the investigation of missing data, univariate normality, multivariate normality, identification of univariate outliers and multivariate outliers. Descriptive statistics showed that missing data did not exist in the dataset. Values of the skewness of the 28 items ranged from .051 to. 264, and values of their kurtosis were between-0.897and -0.397. With reference to the cut-off criteria of the skewness (+/-3) and the kurtosis (+/7)

for univariate normality (Kline, 2011, 2016), datasets of the 28 items were normally distributed. Moreover, visual inspection of the histograms with normality curves, box plots, and Q-Q plots indicated approximate normal distribution. In normality examination with box plots, four univariate outliers were identified (Case 138, Case 167, and Case 175 and case 233) and were excluded.

Following the deletion of the univariate outliers was the detection of multivariate outliers. In line with the critical Chi-square value of 56. 892 (α = .001, *df* = 28), the values of the Malhalanobis distance of 24 cases were greater than the cut-off point, hence were removed from the model. The removal reduced the sample size to 218, meeting the suggested requirement: The sample size of greater than 200 is considered as rather large sample size for the CFA. The subsequent regression analysis revealed that the values of Tolerance of the 28 items were above the cut-off value of .2, and the values of their VIF fell within the acceptable boundary ($\leqslant$ 5), indicating the absence of multicollinearity. However, collenarity and homoscedasticity testing showed that there was bivariate non-normality in the variables; hence, the comprehensive multivariate normality was violated (Beavers et al., 2013; Brown, 2015; Byrne, 2016; Kline, 2016; Reinard, 2006, Rong, 2011*).*

5.2.2.2 The CFA

Examination of offending estimates

The examination of the offending estimates was to ensure the feasibility and the statistical significance of all the parameters estimated. It was a fundamental step before model fit evaluation, which included the inspection of the correlation between constructs (convergent validity), standardized factor loadings and standard errors (Brown, 2015; Byrne, 2016; Kline, 2016; Reinard, 2006, Rong, 2011). According to Brown (2015), Byrne (2016) and Rong (2011), values of correlation coefficients between constructs are supposed to be less than .80, values of standardised factor loadings cannot be close to or exceed 1, and the standard errors should be greater than 0. The statistics of the correlation coefficients, the standardised regression weights and the variances after the first round of the CFA showed that all these parameters were not offending estimates, though correlation coefficient between monitoring and evaluating was .81, slightly greater than .80. Such results suggested that it was appropriate to begin model fit

evaluation. Figure 5.2 displays the correlation coefficients of the factors in Model A before model evaluation.

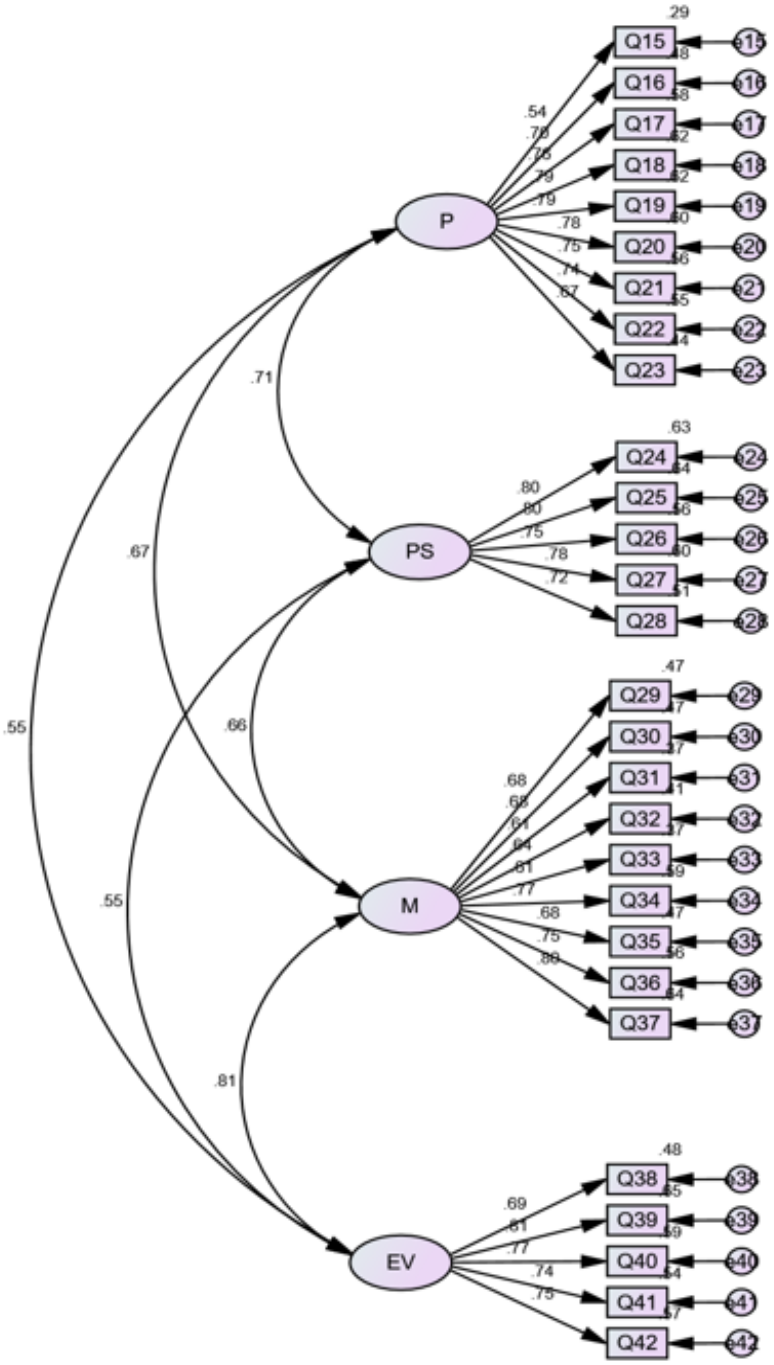

Figure 5.2 *Correlation Coefficients of the Four Factors*

Note. P = planning; PS = problem- solving; M = monitoring; EV= evaluation.

Model evaluation

As multivariate normality was violated as reported above, multivariate normality was re-investigated during the first round of the CFA. The value of the Mardia's coefficient multivariate kurtosis of the dataset was found to be 136.091, and its critical ratio or C.R. was 24.286, both greater than the threshold criteria: Normalised multivariate kurtosis is supposed to be less than 5, and the value of C.R. should be less than 1.96. Therefore, multivariate non-normality was identified. For non-normal correction, bootstrapping procedure was run so that the bias-corrected confidence intervals of the parameter estimate, and the corrected general model fit indices were referred to for model evaluation (Brown, 2015; Byrne, 2016; Kline, 2016; Lewis, 2017; Reinard, 2006, Rong, 2011).

Results of the general model fit indices of Model A were: $\chi 2$ (df = 344) = 750. 034, p = .000. As the value of $\chi 2/df$ was 2.18, larger than the cut-off point (≥ 2), and p value was found to be .00, less than the thumb-up value of .05, the model was supposed to be not satisfactory. Additionally, values of CFI, GFI and TLI were all less than .9, the criteria for an acceptable model, though RMESA and SRMR were less than .08, meeting the thumb-up rule. Given that these indices were estimated under the condition of multivariate non-normality, bootstrap standard errors of each parameters and bootstrap confidence were inspected for bias corrected parameters. After the bias correction, all these indices were statistically significant: p values of the bootstrap standard errors were less than .001, while their bootstrap confidence did not fall on the value of zero. Bollen-Stine bootstrap value was also examined for the bias-corrected general model fit which was equal to zero. The outcome of the bootstrapping was consistent with the original model fit examination, suggesting that Model A did not fit the current dataset and thereby modification was needed for a better model fit (Brown, 2015; Byrne, 2016; Kim, & Millsap, 2014; Kline, 2016).

Model modifications

Model modification was conducted in accordance with factor loadings, modification indices and standardised residual weights. According to Brown (2015), Byrne, (2016) and Kline (2016), acceptable factor loading should be at least greater than .5 and an ideal factor loading should be greater than .7; the observed variables with standardised residual weight greater than 1. 96 for $p < .05$ may indicate areas of strain and should be

removed from the model. In line with this, Q15 and Q33 were deleted because their factor loadings were .54 and .61 respectively, a little greater than .5, but not close to the ideal value of .7. The removal improved the model fit and generated Model B. The inspection of the modification indices of Model B led to the inclusion of extra six paths between error terms (e17 and e18, e24 and e25, e27 and e28, e29 and e30, e31 and e35, e38 and e39), which resulted in a better Model C. Final modification involved the deletion of variables with undesired standardised residual weights (Q19). After the modification, Model D was established with desired model fit indices: Although the index of CFI (.892) was still less than the cut-off value of .9, other indices were rather satisfactory. In addition, the bootstrap estimates proved that the bias-corrected bootstrap standard errors and the intervals of the parameters in the model were all acceptable. In addition, a bias-corrected *p* value of Model D was .204, much greater than the threshold (0.05), indicating the statistical significance of the model fit. Detailed model indices of the four models generated in the CFA are summarised in Table 5.5.

Table 5.5 *Model Fit Indices for Four Rounds of Modifications*

Models	$\chi 2$	CMIN/DF	CFI	GFI	TLI	RMESEA	SRMR
Model A	750.034	2.18	.884	.799	.872	.074	.0616
Model B	628.765	2.14	.897	.814	.886	.073	.0608
Model C	553.975	1.93	.919	.839	.908	.066	.0567
Model D	302.577	1.388	.968	.892	.963	.043	.0512

Model D specified four factors and 23 items. Figure 5.3 illustrates the factor loadings of the 23 items and the correlation coefficients of the four factors.

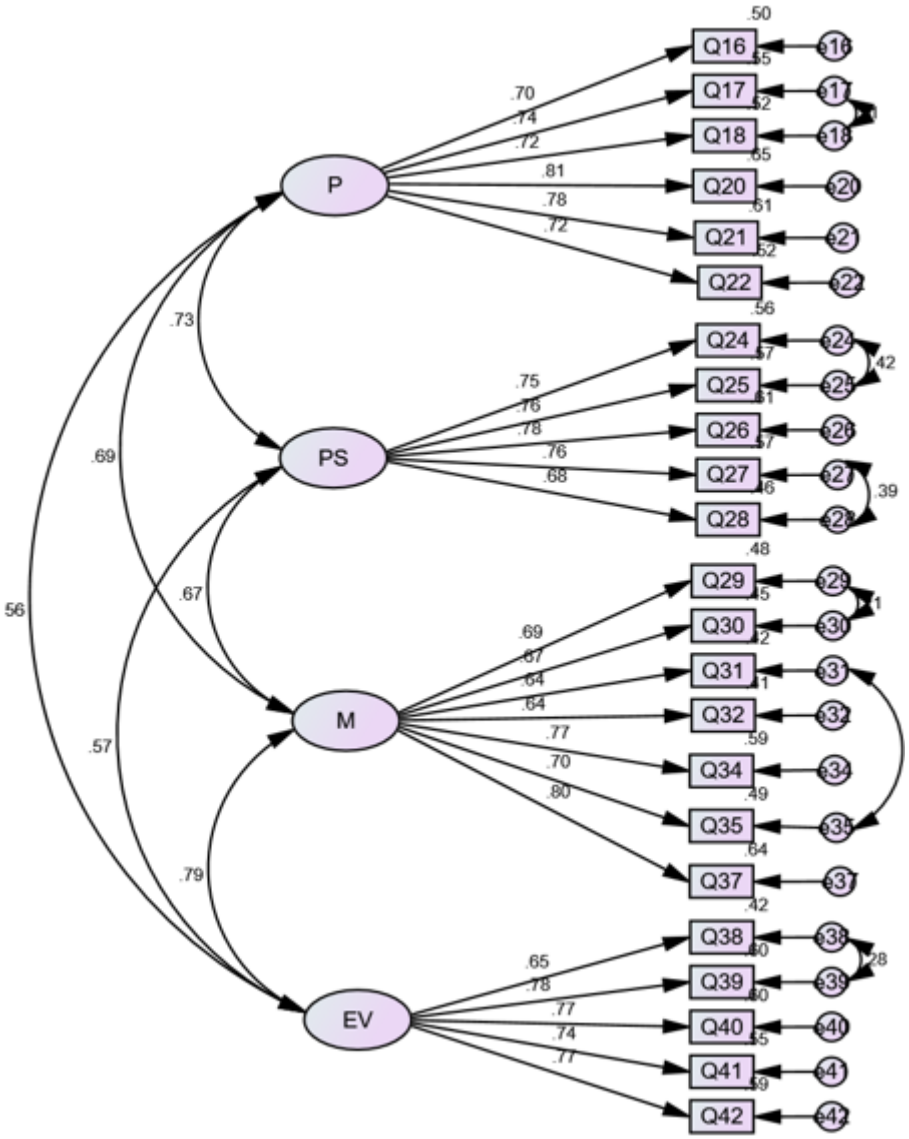

Figure 5.3 *Model D*

Note. P = planning; PS = problem- solving; M = monitoring; EV = evaluation.

Validity and reliability

As Hair (2019) and Hair et al. (2014) proposed, when measuring the fitness of a model in the CFA, it is necessary to assess the validity and reliability of the model with reference to the value of the Composite Reliability (CR > 0.7), Average Variance Extracted (AVE > 0.5) and Maximum Shared Variance (MSV ≤ AVE). As shown in Table 5.6, the values of CR for all the factors were satisfactory; the value of AVE for the factor of monitoring was a little less than the cut-off criterion of .5; and the values of MSV for factors of monitoring and evaluation were slightly greater than the values of their AVE. Those numbers indicated that Model D did not meet the requirements on

construct validity. However, given the fairly large sample size, and the ideal overall fit indices as demonstrated by Table 5.5, this minor discrepancy between the actual values and the cut-off points was tolerable (Brown, 2015; Byrne, 2016; Kline, 2016; Reinard, 2006; Rong, 2011). Taken together, Model D was accepted to fit the dataset.

Table 5.6 *Validity and Reliability of Model D*

Factors	**CR**	**AVE**	**MSV**
M	0.893	0.483	0.664
P	0.910	0.558	0.452
PS	0.878	0.591	0.436
E	0.867	0.566	0.664

Note. M = Monitoring; P = planning; PS = Problem- solving; E = evaluating.

5.2.3 Summary

As was hypothesised, the administration of the EFA on the first sample (N = 254) generated a four-factor model for the MSI, while the employment of the CFA on a separate sample (N = 242) validated the model with robust reliability and validity. The results validated the use of the MSI as one of the key instruments to elicit the Chinese EFL learners' metacognitive strategy use in performing the integrated speaking test tasks with varying levels of complexity.

In accordance with Table 3.1, the first factor, linguistically labelled as planning, refers to the Chinese EFL learners' determination on their objectives and how to achieve the expected goals in test performance. This planning construct is reflected by six items (e.g. 我清楚题目要求我做什么； 我明白需要规划答题过程). The second factor is labelled as problem-solving which is highly related to what the Chinese EFL learners did when encountering problems in performing test tasks such as making a guess or using substitute. Five items represent this factor and examples of these items include 我会借助已有知识(如上单词的下文，构词及话题) 来猜测陌生单词或词组的意思. The third construct is monitoring, which is an examining process in performing test

tasks for a given plan. Items in Model D for this construct are: (a) 我知道答题中何时要更加仔细; (b) 口试过程中，我知道自己哪些地方说得比较地道. The fourth factor is evaluating, which is to investigate the Chinese EFL students' response to post-test self-evaluation. This construct is represented by items such as (a) 我会评价自己在答题表现方面的满意; and (b) 任务结束后我检查自己是否达到目标. Table 5.7 provides the detailed components of Model D.

Table 5.7 *Factors and Items in Model D*

Factors	**Sub- constructs**	**Items**
P	Setting Goals	我清楚题目要求我做什么。
		我确信已清楚任务目标。
	Self-Management	我明白需要规划答题过程。
		我明白完成任务所需的主要步骤。
	Organizational Planning	我想过需要做什么才能完成好答题。
		我事先组织好了想说的内容的结构。
PS	Inference	我会借助已有知识(如上单词的下文，构词及话题)来猜测陌生单词或词组的意思。
		我根据上下文猜测话题。
		我借助已有知识来完成话题任务。
	Substitute	(口语表达)想不起来某个英语单词时，我用相通常思的其他词或词组。
		我使用近义词之类其他方法来表达意思。
M	Selective attention	我知道答题时何时该加快速度。
	Selective attention	我知道答题中何时要更加仔细。
	Ask if it makes sense	答题过程中我知道自己用掉多少时间。
	Ask if it makes sense	口试过程中，我知道自己哪些地方说得比较地道。
	Personalize/personal experience	我能把题中的信息与己有知识联系起来。
	Take notes	答题过程中我会记录重要词语与概念。
E	Evaluating performance	完成口语任务后我在脑子里给自己的表现打了个分数。
		我会评价自己在答题表现方面的满意度。
	Check goals	任务结束后我检查自己是否达到目标。
		我会评价原定计划实施的有效性。
	Verify predictions and guesses	考试后我会检查自己的错误。

Note. M = Monitoring; P = planning; PS = Problem-solving; E= evaluating.

Furthermore, the rather high correlation coefficients of the four constructs validated by the CFA (see Figure 5.3) demonstrated the interaction of the four individual metacognitive strategies in the Chinese EFL learners' response to testing tasks, which was consistent with the working mode of the construct in this study: Metacognitive strategies operate either independently or interactively in a cluster (see Sections 3.2.4).

5.3 Phase Two: Answering Research Questions

Given that the complexity of the statistical analyses in Phase Two might pose challenges to smooth reading, a brief summary of the key quantitative results was presented.

a) Task difficulty perceived by the participants had a significant correlation with mental efforts and varied across tasks. The results indicated that task difficulty could serve as an indicator of task complexity in the actual statistical analyses, and task complexity varied across the four integrated speaking test tasks.

b) The Chinese EFL learners' metacognitive strategy use (either in a clustering form or in an individual manner) had no substantial effects on their speaking performance.

c) Task complexity had a significant effect on the Chinese EFL learners' reported use of clustering metacognitive strategies. Except for problem-solving, no strong association were identified between task complexity and the individual metacognitive strategies.

d) Significant variance was discovered in the Chinese EFL learners' speaking performance across task complexity.

e) The Chinese EFL learners' reported use of the individual metacognitive monitoring had a significant moderating effect on the relationship between task complexity and their speaking performance. Similar moderating effects were identified in some interactive metacognitive strategies when they were investigated simultaneously with the four individual metacognitive strategies.

5.3.1 Step One

Quantitative results in Step One had to do with the fundamental information on the Chinese EFL learner's metacognitive strategy use, their oral scores and the assumed variability of task complexity across tasks. The results were used to address Research Question 1, Research Question 2, and Research Question 4.

5.3.1.1 Individual metacognitive strategies across tasks

Research Question 1 related to the metacognitive strategies reported by the Chinese EFL learners across the integrated speaking test tasks. Descriptive statistics (see Table 5.8) revealed that the means of all the four reported individual metacognitive strategies fell in the range from 3 to 4. With reference to the MSI in which "3" stands for "often" and "4" represents "usually" on the six-point Likert scale, this range value indicated that the students were moderate users of metacognitive strategies across the four integrated speaking test tasks. To be specific, problem-solving was used the most followed by planning and evaluating, whereas monitoring was the least frequently used strategy.

Table 5. 8 *Means of Individual Metacognitive Strategies across Tasks*

Tasks	**Planning**	**Problem solving**	**Monitoring**	**Evaluating**
Task 3	3.605	3.899	3.162	3.168
Task 4	3.376	3.445	3.205	3.220
Task 5	3.525	3.688	3.176	3.363
Task 6	3.546	3.740	3.300	3.260

5.3.1.2 Task complexity across tasks

Research Question 2 regarded the variability in task complexity across the four integrated speaking test tasks. Task complexity was statistically measured via the students' perceptions of task difficulty after a strong correlation between task difficulty and mental efforts were found (Révész et al., 2016; Robinson, 2001; Sasayama, 2015, 2016).

Correlation between task difficulty and mental efforts

Results of the Pearson correlation test in Table 5.9 showed that mental efforts had a strong correlation with task difficulty across tasks. This result indicated that task

difficulty could be used as an indicator of task complexity statistically (Révész et al., 2016).

Table 5.9 *Correlation between Task difficulty and Mental Efforts*

Statistics	**Task 3**	**Task 4**	**Task 5**	**Task 6**
Pearson Correlation	.760	.684	.658	.755
Sig. (2-tailed)	.000	.000	.000	.000

Descriptive statistics of task Difficulty

As presented in Table 5.10, the means of the students' ratings on the four tasks were all far above 4.5, ranging from 5.57 to 6.34. Since the number of 9 on the scale referred to extreme difficult tasks, these values suggested that the students perceived the four integrated speaking test tasks as very difficult. In addition, the values showed that Task 5 was evaluated by the students as the most cognitively demanding task followed by Task 6. Task 4 ranked the third and Task 3 was rated as the easiest one. Moreover, values of the skewness ranged from .070 to .076, and those of the kurtosis were between -. 615 and -. 373, all meeting the required threshold of normality ($-3 \leq$ skewness ≤ 3; $-8 \leq$ kurtosis ≤ 8) (Filed, 2018; Muijs, 2011; Pallant, 2016).

Table 5.10 *Descriptive Statistics of Task difficulty (Students)*

Tasks	**Means**	**Standard Deviation**	**Skewness**	**Kurtosis**
Task 3	5.57	.076	.097	-.570
Task 4	6.07	.072	-.115	-.551
Task 5	6.51	.070	-.419	-.373
Task 6	6.34	.076	-.387	-.612

Although the descriptive statistics demonstrated that task difficulty of the four speaking test tasks varied, whether significant variability existed was not identified at this stage. To investigate the possible significant variations in tasks difficulty across tasks, a one-way repeated measures ANOVA was conducted (Field, 2018; Frey, 2018; Rahimi, 2016; Pallant, 2016).

One-way repeated measures ANOVA

As task difficulty was measured on a nine-point Likert type scale, it was treated as a continuous variable. The four integrated speaking test tasks indicated four levels of treatment conditions, and therefore were regarded as the independent variable (Byrne, 2004; Frey, 2018; Hancock & Mueller, 2013 Pallant, 2016). Visual inspection of the box plots identified 11 outliers, so the final valid sample size was 605, meeting the cut-off criteria for the subsequent ANOVA: At least 15 participants for each treatment condition (Frey, 2018; Nishishiba, Jones & Kraner, 2014; Pallant, 2016). After outlier deletion, histograms, Q-Q plots, box plots were checked, which further evidenced the approximately normal distribution of the dataset. Results of the Mauchly's test revealed that the assumption of Sphericity was violated [F (5) =116.904, p = .000]. Under such situations, the value of Green-house-Geisser epsilon was referred to for correction (Filed, 2018; Frey, 2018; Muijs, 2011; Pallant, 2016).

Results of the one-way repeated measures ANOVA with Greenhouse-Geisser correction showed that large effects of variance were found in the within-participants factor of the four test tasks on task difficulty [F (2. 646, 1586. 36) = 81.121, p <.001; (η^2) =.119]. Multivariate test revealed that the value of Partical Eta Squared (η^2) was .257, indicating a large effect size. The effect size confirmed the significant influence of task complexity on task difficulty, suggesting that task difficulty varied across the four tasks, which further revealed that values of task difficulty rated by the students could be used to statistically to indicate task complexity (Frey, 2018; Field, 2005; Pallant, 2016).

Expert judgment: Teachers' perspectives

Responses from the English teachers to the open-ended questionnaires showed that they rated Task 5 (Means = 6.25) as the most difficult task, followed by Task 6 (Means = 5), Task 4 (Means = 4.6) and Task 3 (Means = 4) as illustrated in Table 5.11. The teachers' ratings corresponded to those of the students.

Table 5.11 *Teachers' Ratings on Task Difficulty*

Teachers	**Task Difficulty**			
	Task 3	**Task 4**	**Task 5**	**Task 6**
Teacher A	3	5	6	7
Teacher B	2	2	5	2
Teacher C	5	6	7	7
Teacher D	6	5	8	5
Teacher E	4	5	5	4
Means	4	4.6	6.25	5

5.3.1.3 Speaking performance across tasks

Research Question 4 concerned the students' speaking performance reflected by their scores. To ensure the validity of the scores, inter-rater reliability was examined with reference to the Cronbach's alpha coefficient. The index was .911, above the rule of thumb-up (＞.70), indicating the statistical validity of the rated scores (Ergai et al., 2016; Frey, 2018; Hallgren, 2012; Sun, 2016; Upton & Cook, 2014).

Table 5.12 *Descriptive Analysis of Oral Scores across Tasks*

Oral Scores	**Task3**	**Task4**	**Task5**	**Task6**
Means	5.45	4.40	3.51	4.86
Skewness	.151	389	.798	.387
Kurtosis	-.528	-.689	-.211	-.506
SD (Skewness)	.247	.247	.247	.247
SD (Kurtosis)	.490	.490	.490	.490
Z-value (Skewness)	.611	1.574	.43	1.56
Z-value (Kurtosis)	-1.1	-1.4	3.23	1.03

As displayed in Table 5.12, regarding the means of the scores, Task 3 had the highest value, followed by Task 6 and Task 4, whilst Task 5 ranked the lowest. In addition, descriptive analysis revealed no outliers, and the values of the skewness, the kurtosis,

and the Z-values of oral scores on Task 3, Task 4 and Task 6 evidenced normally distribution except Task 5.

5.3.2 Step Two

Quantitative results in Step Two regard the relationships among the three research variables to address Research Question 5, Research Question 6, and Research Question 7 and Research Question 8.

5.3.2.1 Task complexity and speaking performance

Research Question 6 examined the correlations between task complexity and the Chinese EFL learners' oral scores. In the assumption test, after data screening, the valid sample size was 95, meeting the threshold for ANOVA (Field, 2018; Frey, 2018; Muijs, 2011; Pallant, 2016). Descriptive analysis of test scores shown by Table 5.12 revealed that the assumptions were met. In addition, as displayed by Table 5.13, values of Kolmogorov-Smirmov indicated that scores on Task 4 (p = .035) and Task 5 (p = .000) were not perfectly normally distributed because the p values of the students' oral scores on the two tasks were less than the cut-off criteria of .05.

Table 5.13 *Test of Normality*

		Kolmogorov-Smirnova	
Oral Scores	**Statistics**	***df***	**Sig.**
Task 3	.072	95	.200*
Task 4	.095	95	.035
Task 5	.133	95	.000
Task 6	.076	95	.200*

*$p > .05$

However, Pallant (2016) has pointed out that researchers should not rely only on Kolmogorov-Smirmov in assumption tests for ANOVA. Rather, the scholar suggests that histograms, normal Q-Q plots and box plots should be inspected for normality. Visual inspection of the above-mentioned graphs suggested that the dataset of the students' oral scores on Task 4 and Task 5 were approximately normally distributed. Additionally, Mauchly's Test of sphericity showed that the assumption of homogeneity was not violated with F (5) = 6.78, and p = .238 (＞.05) (Allen, 2017; Frey, 2018;

Pallant, 2016). These indexes indicated that assumption of normality was met in the current oral score dataset for the ANOVA.

Following the same procedure as was in measuring the variability of task complexity (see Section 5.3.1.2), another round of ANOVA was conducted and the output of this this round of ANOVA procedure demonstrated that there was a statistically significant difference in the effect of task complexity on the students' oral scores across tasks: [$F(3, 282) = 23.01$, $p < .05$, $\eta^2 = .197$). Since task complexity was investigated as a combination of the four task complexity factors (see Table 3.3), the result implied that the four factors had synergetic effects on the students' speaking performance across tasks.

5.3.2.2 Metacognitive strategy use and speaking performance

Research Question 5 was on the correlations between the metacognitive strategies reported by the Chinese EFL learners and their speaking performance. Assumption testing for a multiple linear regression procedure to address the research question includes the examination of sample size, normality of the dependent variable, multicollinearity, outliers, normality, linearity, homoscedasticity, and independence of residuals.

Data preparation yielded 95 participant samples. It met the sample size requirement on multiple linear regression testing: N = 50 + 8M (M means the number of independent variables or the four individual metacognitive strategies in this study) (Frey, 2018). Data on students' test scores were approximately normally distributed as illustrated in Table 5.13. Indices of correlations within the four metacognitive strategies across the four tasks fell in the range of -.027 to .731, all less than .08, suggesting no multicollinearity. The standard residuals of the oral scores on all the tasks were between -3 and 3, within the cut-off value for linearity. Values of the Cook's distance for the four tasks were far from 1, indicating the absence of outliers. Q-Q plots for each task's model revealed the normal distribution of the score residuals. Scatter plots, in the meanwhile, suggested no homoscedasticity of the score dataset across the four tasks (Allen, 2017; Frey, 2018; Pallant, 2016).

Results of the multiple linear regression analysis showed that there were no significant collective and independent effects of the metacognitive strategies reported by the students on their oral scores across tasks.

Table 5.14 *Relationship between the Clustering Metacognitive Strategies and Speaking Performance across Tasks*

Tasks	**Adjusted R^2**	***df***	**F**	**Sig.**
Task 3	-.36	4	.18	.95
Task 4	-.00	4	.86	.49
Task 5	.011	4	1.27	.29
Task 6	.008	4	1.19	.32

As shown in Table 5.14, values of the adjusted R^2 on the four tasks were less than .01, suggesting poor model fit. Alternatively stated, the four clustering metacognitive strategies explained a little in the variance of the oral scores across the tasks. In addition, the *p* values of the four tasks were all larger than .05, indicating that the four models built on the dataset of the four tasks were not significant predictors of the students' speaking performance across tasks. The results implied that no substantial effects of the clustering metacognitive strategies on the students' speaking performance across tasks were discovered.

Further, as presented in Table 5.15, except for problem-solving used on Task 5, all the *p* values of the *B* coefficients for the four subcomponents of the metacognitive strategies on the four test tasks were larger than .05. Such results revealed that generally the four individual metacognitive strategies had no significant effects on the students' speaking performance across tasks.

Table 5.15 *Relationships between Individual Metacognitive Strategies and Speaking Performance across Tasks*

Tasks	**Metacognitive Strategies**	***B***	**t**	**Sig.**
Task 3	Planning	.059	.219	.827
	Problem- solving	.257	.646	.520
	Evaluating	.052	.131	.896
	Monitoring	-.112	-.270	.788
Task 4	Planning	.106	.672	.503
	Problem- solving	.032	.287	.774
	Evaluating	.153	1.105	.272
	Monitoring	-.085	-.530	597
Task 5	Planning	-.019	-.036	.972
	Problem- solving	.946	2.158	.034
	Evaluating	-.711	-1.221	.225
	Monitoring	.185	.413	.681
Task 6	Planning	.903	1.831	.070
	Problem- solving	-.274	-.545	.587
	Evaluating	.304	.657	.513
	Monitoring	-.623	-1.263	.210

5.3.2.3 Metacognitive strategy use and task complexity

Research Question 7 had to do with the relationships between metacognitive strategy use and task complexity across tasks. The normality examination of the values of the skewness and the kurtosis revealed that dataset of the four individual metacognitive strategies met the cut-off value ($-3 \leq$ skewness ≤ 3; $-8 \leq$ kurtosis ≤ 8) for normal distribution. Visual inspection of the Q-Q plots, histograms and box plots also evidenced approximate normality. Four univariate outliers and two multivariate outliers were found with reference to the critical chi-square value ($df = 4$, $\alpha = .001$) which was 18.647. The exclusion of the outliers reduced the sample size to 95, but still met the rule of thumb for MANOVA. Matrix of scatterplots showed the existence of linearity in the four dependent variables. The subsequent bivariate correlation test suggested that the strength of the correlations between variables were between .490 and .697, less than .08, indicating desired multicollineartiy and singularity. Box's test via regression showed that *p* value was less than .001, suggesting homogeneity violation. In spite of this, Leven's test proved that the *p* values of the four dependent variables were larger than .25, indicated that the assumption of equality of variance of the variables was not violated. Given the assumption test results, indices of Pillai's trace were used for correction in the MANOVA test (Allen, 2017; Field, 2018; Pallant, 2016).

The more robust Pillai's trace indices pointed out that there were significant within-subject difference across task complexity on the combined dependent variables or the students' reported use of the clustering metacognitive strategies: F(12, 1212) = 12, p = .007 (less than the threshold of .05), and Partical Eta Squared (η^2) = .022. The result demonstrated significant difference in the synergetic effect of task complexity on the clustering metacognitive strategies in the Chinese EFL learners' performance across tasks.

To further locate the diffidence in the four individual metacognitive strategies across tasks, a series of separate ANOVAs were conducted. Each ANOVA was evaluated at an alpha level of .25 with Bonferroni adjustment (Pallant, 2016). Results displayed that the students' reported use of problem-solving demonstrated significant heterogeneity across tasks [F (3, 405) = 3.853, p = .010, η^2 = .022], whilst significant variations were not found in the other three metacognitive strategies: Planning [F(3, 405) = 1. 205, p = .381, η^2 = .008]; Monitoring [F (3, 405) = .415, p = .743, η^2 = .003]; and Evaluating [F (3, 405) = .730, p = .474, η^2 = .006]] (Frey, 2018; Field, 2005; Muijs, 2011; Pallant, 2016).

5.3.2.4 Metacognitive strategy use, task complexity and speaking performance

Research Question 8 investigated the complicated relationships among the three research variables in two research steps as described in Section 4.7.2. In fact, except for the predictor variables at Level-1, data analyses in the two steps followed the same procedure. Due to this, report of the relationships among the three variables mainly focused on Step One.

Assumption testing

After outlier examination in line with Level-1 residuals, sample size for the HLM was 95, meeting the threshold of at least k (independent variables) + 2 observations at Level-1 and 60 upward at Level-2. Results of homogeneity test of Level-1 variance revealed that F (df = 77) = 10.57882, and p value >.500, indicating homogeneity violation in Level-1 residuals, therefore, model building was administered in the setting of heterogeneity with reference to robust statistics. Visual inspection of the scatter plots, O-Q plots and histograms of Level-1 and Level-2 residuals showed that there were no multicollinearity, heteroscedasticity, and violation of residual normality. Hence, assumption was met on the current dataset for the HLM.

Feasibility testing

Following assumption testing was the feasibility testing to investigate the appropriateness of the HLM in the current dataset. For this purpose, between-subject variations in the students' oral scores and their metacognitive strategy use across task complexity were inspected, multicollinearity detection in Level-2 predictor variables was administered, and the value of the ICC of the null model on the HLM was examined (Barkaoui, 2010, 2013; Garson, 2013; Raudenbush & Bryk, 2002; Snijders, 2010).

Before the null model was run, the examination of between-subject variance in oral scores and metacognitive strategy use was achieved via one-way ANOVA and box-whisker plots. The results of the ANOVA showed significant between-subject variability across task complexity [F (df = 3) = 7.365, $p < .001$]. Visual inspection of the box-whisker plots pointed to variability in the students' metacognitive strategy use across tasks. Such between-subject variations suggested the feasibility of the HLM testing on the current dataset (Raudenbush & Bryk, 2002; Weng, 2009). The examination of multicollinearity on Level-2 predictor variables before model building with reference to the cut-off point of VIF (≤ 10) revealed that there was no multicollinearity (see Table 5.16). The result also supported the appropriateness of the HLM testing (Garson, 2013).

Table 5.16 *Multicollinearity Detection on Level-2 Predictors*

P	**Collinearity**	**PS**	**Collinearity**	**M**	**Collinearity**	**E**	**Collinearity**
	VIF		VIF		VIF		VIF
PS	1. 256	P	1. 270	P	1. 418	P	1. 127
M	1. 147	M	1. 154	PS	1. 411	PS	1. 354
E	1. 290	E	1. 291	E	1. 327	M	1. 341

Note. P = planning; PS = problem- solving; M = monitoring; E = evaluating.

Model 1 (null model)

The value of ICC is another index for the examination of the appropriateness of the HLM. It was calculated via Equation 1 as presented below in the null model (the unconditional model) condition on the HLM (Barkaoui, 2010, 2013; Garson, 2013; see also, Huta, 2014; Meissel, 2014; Nett et al., 2012; Nezlek, 2012; Wen, 2009; Weng, & Qiu 2014; Raudenbush & Bryk, 2002; Snijders, 2010).

$\rho = \tau_{00} / (\sigma^2 + \tau_{00})$ Equation 1

In Equation 1, ρ is the value of the ICC, τ_{00} refers to Level-2 variance (the random effects of the students' differences on their oral scores); and σ^2 refers to Level-1 variance (the random effects of intra-individual variability of the students on their oral scores). As shown by Table 5.17, τ_{00} for the current data was 5.66, while σ^2 was 3. 41. In accordance with Equation 1, ICC= 5.66 / (5.66 + 3.41) = 0. 622, indicating that 62 % of the total variance in the students' oral scores was explained by the students' individual differences at Level-2, whilst task complexity explained about 38% of the total variance in oral scores. The result, while further indicating the necessity and appropriateness of the HLM, cross-validated the ANOVA outcome for Research Question 6 reported above. Moreover, as described earlier, the null model reported the benchmark value of the deviance for model comparison. As seen from Table 5. 17, the deviance value of the null model was 1737.861. Taken together, the results of the above feasibility testing suggested that the HLM suited the present study.

Table 5.17 *Model 1(Null Model)*

Random Effect	**Variance component**	**SE**	***p***
Level 2 MOS across tasks (π_{0i})	5. 65523	2. 37807	< 0.001
Level 1 Unexplained variance (e_{ti})	3. 41255	1. 84731	
Deviance		1737. 861	

Note. MOS = means of student's oral scores; SE ==Standard error; the number estimated parameters for deviance = 15; *p < .05; ** p < .01.

Model 2 (full model)

Model 2 was established to examine the relationships among the three research variables for Research Question 8. Equations of the two levels in the model were presented in the followings:

Level-1 Model: $SCORES_{ti} = \pi_{0i} + \pi_{1i} * (TASK\_DIF_{ti}) + e_{ti}$

Level-2 Model: $\pi_{0i} = \beta_{00} + \beta_{01} * (\text{PLANNING}_i) + \beta_{02} * (\text{PROBLEM}_i) + \beta_{03} * (\text{MONITORI}_i) + \beta_{04} * (\text{EVALUATI}_i) + r_{0i}$; $\pi_{1i} = \beta_{10} + \beta_{11} * (\text{PLANNING}_i) + \beta_{12} * (\text{PROBLEM}_i) + \beta_{13} * (\text{MONITOR}I_i) + \beta_{14} * (\text{EVALUATI}_i) + r_{1i}$

Mixed Model: $\text{SCORES}_{ti} = \beta_{00} + \beta_{01} * \text{PLANNING}_i + \beta_{02} * \text{PROBLEM}_i + \beta_{03} * \text{MONITORI}_i + \beta_{04} * \text{EVALUATI}_i + \beta_{10} * \text{TASK\_DIF}_{ti} + \beta_{11} * \text{PLANNING}_i * \text{TASK\_DIF}_{ti} + \beta_{12} * \text{PROBLEM}_i * \text{TASK\_DIF}_{ti} + \beta_{13} * \text{MONITORI}_i * \text{TASK\_DIF}_{ti} + \beta_{14} * \text{EVALUATI}_i * \text{TASK\_DIF}_{ti} + r_{0i} + r_{1i} * \text{TASK\_DIF}_{ti} + e_{ti}$

In these equations, $SCORES_{ti}$ indicated the *i* th student's oral scores on the *t* th task. π_{0i} referred to the average means of the *i* th student's oral scores across the four tasks, which was determined by the average means of the student's oral scores across the four integrated speaking tasks and individuals (β_{00}). π_{0i} was also determined by β_{01}, β_{02}, β_{03}, β_{04}, and r_{0i}. β_{01}, β_{02}, β_{03}, and β_{04} referred to the fixed effect of the four metacognitive strategies reported by the students on the average mean of their oral scores across task conditions, while r_{0i} indicated the random effects of individual heterogeneity on the means of the student' oral scores across tasks that could not be explained by the model. π_{1i} indicated the slope or the relationship between task difficulty and the students' oral scores, which was affected by β_{10}, β_{11}, β_{12}, β_{13}, β_{14}. β_{10} represented the average slope of task difficulty on the students' oral scores, or the average relationship between task difficulty and oral scores. Values of β_{11}, β_{12}, β_{13}, β_{14} provided the values of the effects of the four metacognitive strategies on the relationship between task difficulty and the students' oral scores, or the cross-level interaction between task difficulty and metacognitive strategies. To be specific, β_{11} indicated the impact of planning on the relationship between task difficulty and oral scores; β_{12} was used to investigate the effect of problem-solving on the correlations between Level-1 variables, while β_{13}, β_{14} denoted the effects of monitoring and evaluating on the associations between Level-1 variables respectively. r_{1i} referred to the inter individual variations in the relationship between the task difficulty and the average oral scores that were inexplicable in the model. By the same vein, e_{ti} suggested the random effects of the intra individual heterogeneity in task difficulty on oral scores which could not be accounted for by the model (Barkaoui, 2010, 2013; Garson, 2013; Meissel, 2014; Randenbush & Byrk, 2002; Weng, 2009).

As Research Question 8 focalises on the effects of the interactions between metacognitive strategies and the task difficulty on oral scores, special attention was paid to coefficients of β_{11}, β_{12}, β_{13}, β_{14}, which suggested the cross-level interactions. Table 5.18 gives a display of these coefficients.

Table 5.18 *Cross-Level Effects in Model 2*

Fixed Effect	**Coefficients**	**SE**	***t*-ratio**	***p***
Average relationship between TD and OS (β_{10})	-.205	0.075	-2.74	0.007**
Effects of P on the relationship between TD and OS(β_{11})	0.124	0.149	0.830	0.408
Effects of PS on the relationship between TD and OS(β_{12})	-0.236	0.131	-1.793	0.076
Effects of M on the relationship between TD and OS(β_{13})	0.289	0.136	2.122	0.037*
Effects of E on the relationship between TD and OS (β_{14})	-0.262	0.150	-1.74	0.085

Note. SE = Standard error; TD = task difficulty; OS = students' oral scores; P = Planning; PS = Problem-solving; M= Monitoring; E = Evaluating; *p < .05; **p < .01.

As seen from Table 5.18, p values of β_{11}, β_{12}, β_{14} were .408, .007, and .085 respectively, all greater than .005. It suggested that planning, problem-solving, and evaluating did not have significant effects on the relationship between task difficulty and the students' oral scores. On the other hand, the p value of β_{13} was less than .05 (t-ratio = 2. 122; p = .037), which revealed that monitoring had a substantive impact on the relationship between task difficulty and oral scores. In addition, the value of β_{13} was .289, indicating the magnitude and the direction of such impact: When other factors were controlled, the effect of task difficulty on the oral scores of the students who used monitoring more frequently were weaker, reducing by .289 units, compared with such effect on those who use the strategy less frequently. Simply put, monitoring moderated the negative effect of task complexity on the Chinese EFL learners' speaking performance across the four integrated speaking tests. It implied that the more frequently a Chinese EFL learner used metacognitive monitoring, the weaker effect of task complexity on his or her speaking

performance occurred (Barkaoui, 2010, 2013; Garson, 2013; Raudenbush & Bryk, 2002).

In the meanwhile, the value of β_{10} was -.205, suggesting that the average slope of task difficulty on the students' oral scores across individual was expected to be -.205. The minus value pointed out the negative direction of the relationship between task difficulty and oral scores, implying that the more complex a given task is, the lower the students' test score will be. Furthermore, *p* value of β_{10} was .007, less than .05, which suggested that the average slope of task difficulty on students' oral scores demonstrated heterogeneity. Additionally, *p* values of β_{01}, β_{02}, β_{03}, β_{04} were all greater than .05: β_{01} = -0. 181 (*t* - ratio = -0. 340; SD = 0. 533; *p* = .037), β_{02} = 0.318 (t - ratio = 0. 671; SD = 0. 473; *p* = .504), β_{03} = 0.100 (*t-ratio* = 0.168; SD = 0.568; *p* = 0.867), β_{04} = 0.461 (*t*-ratio = 0.779; SD = 0.591; *p* = .085). These results are summarised in Table 5.19, which suggested that variance in the students' use of planning, problem solving, monitoring and evaluating had no direct effects on their oral scores across tasks. Further, these results added more evidence to the relationships between the Chinese EFL learners' use of metacognitive strategies and their speaking performance examined previously which revealed that metacognitive strategies had no effects on speaking performance in both clustering and individual manners at within-subject levels. Therefore, it can be concluded that the Chinese EFL learners' reported use of metacognitive strategies had no impacts on their speaking performance in both clustering and individual manners at both within-subject and between-subject levels.

Table 5.19 *Fixed Effects in Model 2*

Fixed Effect	**Coefficient**	**SE**	***t*-ratio**	***p***
The relationship between P and OS across tasks (β_{01})	-0.181	0.533	-0.340	0.735
The relationship between PS and OS across tasks (β_{02})	0.318	0.473	0.671	0.504
The relationship between M and OS across tasks (β_{03})	0.100	0.568	0.168	0.867
The relationship between E and OS across tasks (β_{04})	0.461	0.591	0.779	0.438

Note. SE= Standard error; P = planning; OS = students' oral scores; PS = problem-solving; M = monitoring; E = evaluating; *$p < .05$; ** $p < .01$.

In terms of random effects, Table 5.20 illustrates that the value of the variance component for r_{0i} was 5.808 [F (77) = 683.716, SD = 2.41], and its *p* value was less than .01, implying that strong between-subject difference existed in the students' mean oral scores across tasks. In contrast, *p* value of r_{1i} (F (77) = 93.616, SD = 0. 166] was 0.096, greater than the cut-off value of .05. Despite this, r_{1i} was considered statistically significant as Nezlek (2012) proposed that $p < .10$ was acceptable "…in decisions about the inclusion or exclusion error terms…" (p. 225). Henceforth, the average relationship between task difficulty and the students' oral scores varied across individuals at Level-2. In addition, values of π_{0i} (the reliability estimate for Level-1 intercept) was 0.889, and above .8, suggesting that the variance in the students' mean oral scores across tasks was explicable by the individual-level predictors (not metacognitive strategies in this study) (Barkaoui, 2010, 2013; Garson, 2013; Meissel, 2014; Randenbush & Byrk, 2002; Weng, 2009).

Table 5.20 *Random Effects in Model 2*

Random effect	**SE**	**Variance**	χ^2 **(*df*)**	***P***
MOS across tasks (r_{0i})	2.41	5.808	683.716 (77)	<0.001**
the average relationship between TD and OS (r_{1i})	0.166	0.028	93.616 (77)	0.096*
Random Level-1 coefficient		Reliability estimate		
Intercept (π_{0i}): MOS across tasks		0.889		
Deviance		1672.999		

Note. SE =standard error; MOS = means of students oral scores; TD = task difficulty; OS = students' oral scores; *p < .05; ** p < .01.

Model checking

Model checking involved the examination of model misspecification, and model fit. According to Garson (2013), Nezlek (2008, 2011) and Raudenbush and Bryk (2002), if the difference between ordinary standard errors and robust standard errors in the finial estimation of fixed effects is identified as strong, model misspecification will exist. Visual inspection of the two types of standard errors showed that there was no significant variance; hence, model specification was acceptable. Examination of model fit was conducted with reference to the deviance value via model comparison between the null model and the full model. Figures in Table 5.17 (null model) and in Table 5.20 (full model) showed that the values of deviance reduced considerably from 1737.861 (null model) to1672.999 (full model), indicating a better model fit. Additionally, investigation of residuals based on the residual files at the two levels illustrated that the full model fitted well the current dataset (Barkaoui, 2010, 2013; Garson, 2013; Meissel, 2014; Randenbush & Byrk, 2002; Weng, 2009).

Summary of Step One

Results from the first research step revealed that the Chinese EFL learners' reported use of the individual monitoring had a significant effect on the relationship between task complexity and their speaking performance. In the meanwhile, the results indicated that the variance in the Chinese EFL learners' performance across the four integrated

speaking test tasks was associated with the variability in both the test tasks and the test-takers' attributes (not their use of strategic competence).

Step Two

Following the same procedures in Step One, Step Two explored the interactions between the interactive metacognitive strategies and task complexity, as well as the impacts of the interactions on the Chinese EFL learners' speaking performance. Results showed that when the impacts of the four individual metacognitive strategies on the relationship between task difficulty and oral scores were taken into consideration simultaneously, the cross-level interactions took place in the following situations:

a) Problem-solving interacted with monitoring.

b) Planning, problem-solving and evaluating interacted with one another.

c) Problem-solving, monitoring and evaluating interacted with one another.

d) Planning, monitoring and evaluating interacted with one another.

Table 5.21 illustrates the cross-level interactions where *p* values of the effects of the above interactive strategies on the associations between Level-1 variables were all below than .05, suggesting substantial effects of these interactive metacognitive strategies on the association between task complexity and speaking performance. By contrast, all the *p* values of the fixed effects of the interactive strategies on speaking performance were larger than .05, indicating that these interactive strategies did not exert significant effects on speaking performance across individuals.

Table 5.21 *Effects of the Interactive Metacognitive Strategies on the Relationship between Level-1 Variables*

Interactive MS	**Coefficient**	**Standard Error**	***t*-ratio**	***p***
PS × M	-.192	0.94	-2.038	0.044
P × PS × E	-0.025	0.012	-2.046	0.043
PS× M × E	-0.026	0.010	-2.713	0.008
P × M × E	-0.030	0.012	-2.446	0.017

Note. MS = metacognitive strategies; PS = problem solving; M = monitoring; E = evaluating; *$p < .05$; ** $p < .01$.

Chapter Six

Qualitative Findings

6.1 Chapter Overview

"Understanding comes not from the subject who thinks, but from the other that addresses me. This other…is this voice that awakens one to vigilance, to being questioned in the conversation that we are" (Risser, 1997, p. 208). In this chapter, the voices of the participants are presented, providing insights into the Chinese EFL learners' perceptions of metacognitive strategy use and task complexity, as well as insights into the Chinese EFL teachers' judgements of the task complexity.

The students' responses and the teachers' views were reported separately. Equal attention was given to the two groups with special attention paid to the extreme cases and cases with uniqueness, as they helped to explain the qualitative data from various angles. Results on metacognitive strategy use were reported in sequence from planning, problem-solving, and monitoring to evaluating. Similarly, findings on task complexity were divided into five sections from the overall measurement on complexity, planning time, prior knowledge, and steps involved to task type.

6.2 Students' Strategy Use and Perceptions of Task complexity

The student interviewees' metacognitive strategy use is illustrated in Table 6.1. As shown, responses to the Semi-structured Interview Guide were coded into 14 themes. Although some new themes (theme that could not be generated in accordance with the coding scheme) emerged as presented by Table 6.2, and themes that were assumed to be generated in light of the coding scheme were not identified as shown in Table 6.3, the coding result, in general, aligned with the coding scheme (see Table 4.19).

Table 6.1 *Interviewees' Metacognitive Strategy Use*

Subcategory	**Interviewees**	**Reported use of Metacognitive Strategies**
Planning	All	Organisational planning (Outlining and organization)
	In 4 & In 8	The activation of prior knowledge
	In 2 & In 3	Setting-goals
	In1 & In 3	Variation in setting goals response to task complexity
	In 4, In 5, In 6, In 7, In 8	No particular goals
	In 1 & In 3	Chance of oral practice
Problem-solving	In 6, In 7, In 1 In 2, In 3	Inference
	In 8	The use of make-up
	In 8	The use of mother tongue
	In 8 & In 5	The use of gap fillers
	In 1 & In 7	Not knowing how to solve problems
	In 4	No problem-solving
Monitoring	In 1	Taking notes down in a messy manner
	In 6	Purposefully taking notes
	In 5	Making predictions and judgments in processing tasks
Evaluating	In 7 & In 8	Self-evaluation on task completion
	In 2 & In 6	Self-evaluation on understanding the tasks
	In 7	Self-evaluation on fluency of his speech
	In 1& In 3	Self-evaluation on English learning
	In 5	Self-evaluation on his summary of the tasks
	All	Self-evaluation on personal feelings

Note. In = Interviewee.

Table 6.2

New Themes with Examples

Categories	Subcategories	Themes	Examples
Metacognitive strategies	Planning	No particular goals	I had no any goals, and I just thought about finishing the tasks carefully.
		Chance of oral practice	Well, I just wanted to practice my oral English
		Variance in setting goals based on task complexity	When the tasks became more difficult like Task 5, I felt hard to deal with them. And I had no goals at all. Actually, at the beginning, I had... But when the tasks were not what I expected, I felt stressful and lost all my goals in the end.
	Problem solving	Not knowing how to solve problems	There was so much information, and if I couldn't understand the materials given, I had to ignore them and gave them up.
		No problems encountered	I don't think I had problems. You (the researcher) has explained clearly the task requirements …I was very familiar with the test form and the tasks themselves
Task complexity	Planning	Self-evaluation based on personal feelings	Generally, I didn't feel good today. I believed I could have performed better. I think there must be some space for improvement
		Peer Pressure	While I am doing nothing, my peers and classmates may think actively, and to take notes. In such circumstances, I will be more nervous….
Task complexity	Overall judgment	Familiarity	…I think the words, sentences…are very familiar to me
		Mental fatigue	… I felt too tired to perform it.
		Psychological pressure	Task 5 is completely different from what I predicted… which caused pressure on me.
		Fewer/more information Load	Task 4 requires extra reading, so I had to process extra information

Table 6.3

Unidentified Assumed Themes

Categories	**Subcategories**	**Unidentified Codes**
Metacognitive strategies	Planning	Directed attention
		Self-management
	Monitoring	Self-talk
	Evaluating	Verify predictions and guesses
		Check goals
		Evaluating performance

The integration of the information revealed in Table 6.1, Table 6.2 and Table 6.3 suggest that the interviewees were not active users of metacognitive strategies and their metacognitive strategy use was closely correlated with their individual attributes such as motivation, pressure, and prior experiences and knowledge, which were further influenced by task complexity. With reference to Figure 3.3, the relationships among metacognitive strategy use, individual attributes, and task complexity indicated that the students' individual attributes served as a mediator between their metacognitive strategy use and task complexity. Detailed findings on the students' metacognitive strategy use and their perceptions of task complexity are reported in the subsequent subsections.

6.2.1 Individual metacognitive strategy use

In general, the students did not have a clear and particular goal, although practice and understanding oneself were understood by two of them as setting goals. They had weak awareness of self-monitoring, and tended to evaluate their performance in light of their personal feelings which were aroused by various stimuli such as the understanding of the tasks and the completion of the tasks. By contrast, the students resorted to the problem-solving strategy actively when they faced problems, and in particular, they prioritised inference by making guesses using the contexts or their prior knowledge.

6.2.1.1 Planning

Regarding planning, most of the interviewees expressed the idea that they did not set any clear goals due to personal factors such as self-interest and pressure. For instance, after thinking for a few seconds, Interviewee 4 recounted that he had not set any specific goals:

> En…Setting goals? No, I didn't set any goals…. To be honest, I believe whether one set goals or not relies on whether the tasks are of his or her self-interest. If one believes that performing the tasks brings no interest, he or she is likely to attach no importance to them, and they will accordingly never set any goals. (Interviewee 4/27/06/2018)

Interviewee 4's comments demonstrated his understanding of the importance of self-interest in setting goals. He believed that the benefits he could get from the tasks were a determinant in his actual strategy use and this benefit or self-interest was essentially a kind of motivation generated by tasks. It was obvious that the tasks did not arouse interviewee 4's motivation partially because of task complexity, as his later responses evidenced this when he complained that the tasks were so challengingly demanding, and he did not want to be bothered by the challenge:

> The tasks were a little difficult for me, and they were so challenging. I didn't believe I could deal with them with my language proficiency. So, I did not bother to do the tasks with efforts. (Interviewee 4/27/06/2018)

Another example came from Interviewee 5 who did not set any goals because of his concern about the pressure from goal setting:

> Any specific goals? En…let me think…I had no particular goals. What I was thinking in performing the tasks was to finish the tasks carefully. I never thought about setting goals. You know, if you set a goal and regard the speaking tests as tasks, you will feel pressure, and the pressure may cause negative influence on you (Interviewee 5/28/06/2018).

It seemed that this respondent was worrying about the possible failure in meeting goals, and such worry placed pressure on him. Interviewee 6 echoed Interviewee 5's worry:

> I didn't set any goals. Anyway, I didn't want to give myself any pressure. If I understood the task requirements, I just did it without any goals on mind. (Interviewee 6/28/06/2018)

It is evident that like Interviewee 5, Interviewee 6 believed if he set goals, he would suffer from certain pressure. The two interviewees' accounts seemed to suggest that they regarded setting goals as a pressure producer and managed to avoid such pressure by setting no goals. This understanding of and practice on goal setting might expose their anxiety as language learners facing immediate and possibly complex tasks. In addition to personal attributes, some of which were related to task complexity, the direct and negative effects of task complexity on the interviewees' metacognitive strategy use were also reported. For instance, Interviewee 1 recalled that:

> Compared with the first two speaking tasks, the last two tasks were more difficult with a lot of words that I couldn't understand…I was in a complete loss and got no idea of planning at all. What I felt was a sort of confusion and a lot of messy information in my head. (Interviewee1/ 24/06/2018)

Interviewee 1's experience spoke of the fact that when he faced a complex task, he was unable to use planning. In his explanation, task complexity put him into confusion and loss of ideas. Moreover, task complexity provided him with only the jumbled information which he had difficulty understanding and arranging in performing tasks. In his view, it was task complexity that determined whether he could use planning and whether the strategy could help him do the tasks effectively. Interviewee 3's description of her loss of goals in more complex tasks also evidenced the variability in the student interviewees' planning use in response to task with varying degrees of complexity:

> When tasks became more difficult like Task 5, I felt hard to deal with them. I had no goals at all! Actually, at the beginning, I had goals and I had some expectations of myself. However, when the tasks were too

> complex to be what I expected, I felt stressful and lost all my goals in the end. (Interviewee 3/24/06/2018)

The variation in Interviewee 3's goal-setting exemplified the negative impacts of task complexity on her use of planning. When she found that task complexity was within her ability to cope with, she actively employed the metacognitive strategy of planning. However, when faced with a task that was too complex for her to perform with ease, she experienced stress. Under such stressful conditions, goals were finally lost.

In fact, among the eight interviewees, only two of them (Interviewee 3 and Interviewee 1) reported setting goals. The two students said that they considered performing the test tasks as a chance of oral practice and of understanding their language proficiency, as interviewee 3 responded:

> Well, I just wanted to practice my oral English. Before performing each speaking task, I said to myself that I would make use of my own language to do the tasks and to practice my oral English. (Interviewee 3/24/06/2018)

Similarly, interviewee 1 noted that his goal was to know himself better in English learning:

> En…my goal? En…my goal was possibly to understand my own English proficiency. It turned that that my English was so bad…. And I couldn't handle the speaking tasks as I expected (Interviewee 1/24/06/2018)

Typically, setting goals indicates a possible score range that one tries to get, achievements and improvements in language learning tasks. In this sense, interviewee 3's attitudes implied that she might not have any concrete goals when performing the speaking tasks. Therefore, a chance of oral practice became her seemingly vague goal. In the like manner, it is hard to conclude that interviewee 1 had any goals because of his hesitation in reporting setting goals. Besides, his uncertainty evidenced by his use of the word "possibly" revealed that he was unconfident when talking about his goals. In essence, his goal of understanding his English proficiency only demonstrated his lack of clear goals.

6.2.1.2 Problem-solving

Problem-solving was reported as the most frequently used strategy. Almost all the interviewees had a strong sense of problem-solving, which was well illustrated by interviewee 8 who utilised substitution, one of the subcomponents of problem-solving strategy, in various forms to deal with problems. These included making up new ideas , the use of mother tongue and the use of gap fillers as he reported in the following:

> Occasionally, there was nothing in my head, and I didn't know what to speak, then I began to search in my head something such as words that I could use to fill the gap between sentences like 'well', and 'you know'…In performing tasks, I always thought first in Chinese before speaking in English. If I could completely understand the listening or the reading before speaking, I always made up some ideas immediately after the tasks were given. (Interviewee 8/15/07/2018)

Inference (making educated guesses on contexts or one's prior knowledge), another subcomponent of the problem-solving strategy, was prioritised by the interviewees, as it was the most frequently reported. Furthermore, almost all the interviewees associated the priority of inference to their daily classroom study and practice. Interviewee 7 was of the opinion that inference served as the role of an effective helper as follows:

> When I couldn't understand the listening materials, the only useful way I turned to was to make a guess based on the context or on the reading materials. I didn't rely on the recordings too much, because they were played just once and if I missed something, it was impossible that they would be played twice. (Interviewee 7/02/07/ 2018)

Interviewee 6 shared the same opinion with Interviewee 7:

> The problem I met was the new words that I wasn't able to understand. In this case, I made a guess according to the contexts given, and guessed the meaning of the words and why they were used in this context. This solution worked well with me. (Interviewee 6/28/06/2018)

Despite the active use of problem-solving, Interviewee 7 and Interviewee 1 reported that occasionally they did not know how to solve problems due to the heavy information

load involved in the tasks given. For interviewee 7, he lost his ideas as he did not know how to organise so much information in performing tasks:

> A huge load of information was jammed in my head. Even though I understood the listening and reading materials, I wasn't able to organise all the information in a logic way. So, I could not speak it out in the limited time. (Interviewee 7/02/07/2018)

As Interviewee 7 experienced, Interviewee 1's descriptions of messy ideas in his brain showed that he didn't know how to solve problems, either:

> There was so much information, and if I couldn't understand the materials given, I had to ignore them and gave them up. Anyway, I tried my best and everything seemed messy in that situation. (Interviewee 1/ 24/06/2018)

The two interviewees' experiences implied that when the information provided in the tasks was too much for them, what they got were only messy ideas which could not be used for a logical and organised oral performance. Even worse, when the information load was considered by the interviewees as cognitively challenging above their capability to deal with, it was easy for them to give up just as interviewee 1 did in the actual situation. The two cases revealed that variations in task complexity caused extra information load to some interviewees, and consequently, they had no ideas on how to use the problem-solving strategy to address the problems encountered.

Unlike most of the interviewees who actively used problem-solving, one interviewee reported that he had no problems thanks to his prior knowledge on the tasks given:

> En…I think you (the researcher) has explained very clearly the task requirements both in Chinese and English. To me, I think I've learned the relevant information on the tasks twice. The information made me understand the tasks well, and it helped me a lot, because I was very familiar with the test form and the tasks themselves. (Interviewee 4/27/06/2018)

It can be seen that Interviewee 4's prior knowledge came from his familiarity with the speaking tasks, which rendered him the information needed from every aspect of the tasks including task requirements, task topics and testing procedure. The comprehensive familiarity was the source of his confidence that he had no problems at all. In other words, it was his familiarity with the tasks which reduced task complexity and positively influenced his view of the problems and his use of the problem-solving strategy.

6.2.1.3 Monitoring

As indicated by Table 6.3, the interviewees had a weak sense of monitoring. Only three interviewees reported their use of the strategy. Interviewee 1 and Interviewee 6 talked about taking notes, one subcomponent of monitoring, whilst Interviewee 5 reported his use of several subcomponent strategies of monitoring. Apart from them, monitoring was not mentioned at all by the other interviewees. In the case of Interviewee 1, he reported taking notes, but he thought that these notes were just messy ideas:

> I wrote down some simple ideas that I understood. Well, the notes looked a little messy, but I was able to identify them while I was doing the speaking tasks. (Interviewee 1/24/06/2018)

At first thought, the interviewee's performance of taking messy notes may be questioned as his use of monitoring. The reason is that taking notes is defined as taking down key words and concepts as a subcomponent of monitoring, which is quite different from the interviewee's reported "simple ideas" in "messy notes". Nonetheless, the interviewee's awareness of taking recognisable notes for the subsequent speaking tasks can be seen as an attempt to use monitoring. Differing from Interviewee 1, Interview 6 recounted that he had taken notes purposefully:

> I wrote down some key…some key information. For example, I took notes on some important words that I believe I might use in performing the speaking tasks such as the words on places, time, and suggestions provided by the professor in the lectures. (Interviewee 6/28/06/2018)

When being further asked how he judged the key words and information, Interviewee 6 commented:

> I made the judgment in line of my experiences in taking English tests. My learning experiences also helped me to identify the relevant knowledge that I believed useful in taking the tasks. (Interviewee 6/28/06/2018)

Interviewee 6's purposeful judgment on the key words and information with reference to his previous test-taking and learning experiences in note-taking proved that he was using his prior knowledge and experience to perform the tasks.

In addition to taking notes, several other subcomponents of monitoring were reported by Interviewee 5. Compared with other interviewees, Interviewee 5 seemed to have a better understanding of how to perform tasks, as he said that he knew what he was doing and how to do in the process:

> After the recordings were played, I began to think immediately what the conversations or lectures were about. I wrote the information down, and then I thought about what I was going to speak, how many parts my speech might be divided into, and how I speak. Subsequently, I made judgment according to this analysis. I was very clear about what I was going speak and how to do that. (Interviewee 5/28/06/2018)

It is easy to identify that Interviewee 5 selectively paid attention to key information, periodically examined his understanding of the tasks, made deduction, took effective notes and contextualised his knowledge to finish the tasks. The vivid image generated from the interviewee's personal recounts presented a convincing example of an active user of monitoring in performing tasks. Indeed, both Interviewee 6 and Interviewee 5 had a whole academic year of language training targeting at IELTS, and have passed the examination, which indicates that both of the respondents have a higher level of language proficiency compared with the other interviewees. Their purposeful use of monitoring in tackling tasks seemed to illustrate the close relationship between EFL learners' language proficiency and their use of metacognitive monitoring strategy.

6.2.1.4 Evaluating

All interviewees recounted that they had judged their performance after the speaking tasks, but their reflection implied a message: They evaluated their performance based

on personal feelings. In other words, if they felt good after finishing tasks, they would think that they performed well; otherwise they would believe that they had bad performance. Moreover, most of the interviewees commented that whether they felt good or not in evaluating their test performance was determined, to a large degree, by task complexity. For example, when talking about his negative self-evaluation, Interviewee 1 believed that performing tasks exerted a heavy blow to his confidence in English learning. He further commented the "blow" by linking it to task complexity:

> "I felt my English was completely useless in performing the tasks, especially in performing Task 5! It was so complex, and I felt very stressful. I'm wondering if I had spent years learning English that was not useful in real language use situations!" (Interviewee1/24/06/2018)

His complaint indicated the broken-down of his confidence due to task complexity. As a result of such negative feelings, the interviewee evaluated himself negatively and his confidence in English learning was therefore collapsed. Unlike interviewee 1 who had negative attitude towards learning English in self-evaluation, interviewee 3 saw the gap in her English learning reflected by her bad performance in response to task complexity:

> I thought my performance was extremely terrible! I wasn't able to use the English that I had learned to perform tasks. I believe learning English is useless if you just know some words instead of understanding how to use them. Perhaps, in the future, I will pay more attention to language use rather than focusing on memorising words as I always do in English learning. (Interviewee 3/24/06/2018)

Interviewee 3's self-evaluation revealed that she had learned a lesson from performing tasks for future improvement. Despite this, both Interviewee 1's and Interviewee 3's judgments on their performance were generally negative personal feelings, and these feelings were affected by their perceptions of task complexity.

Unfortunately, apart from evaluating themselves on their personal feelings, interviewees did not report any strategies related to evaluation. For instance, they did not attempt to verify whether their guesses were right or not (one subcomponent of evaluating

strategy), though all of them responded that they actively used guesses or inference to solve problems as reported above.

6.2.2 Perceptions of task complexity

During the interviews on task complexity, "familiarity" was the most frequently mentioned word by the interviewees. Familiarity, according to their interpretation, referred to the presence or absence of the interviewees' prior knowledge on a given task. If they had prior knowledge, they tended to judge that task as easy; otherwise they would comment that the task was complex. Because of this, the interviewees prioritised prior knowledge over other task factors in judging task complexity and regarded it a determinant in examining the roles of planning time, steps involved, and task type in task complexity. Such perceptions of task complexity answered the Research Question 3. In their measurement of the specific individual task complexity factors, consensuses were reached as the followings:

a) The longer planning time a task could render, the less complex the task was;
b) A task type that requires presenting solutions were easier compared with the one that asked task-takers to illustrate an academic concept with examples;
c) Tasks involving three steps (reading, listening and speaking) were less challenging in comparison with tasks involving two steps (listening and speaking) since students treated reading as a prior knowledge provider.

In accordance with the above-mentioned agreements, the student interviewees rated Task 5 and Task 4 as the more complex tasks compared with Task 3 and Task 6. In addition to task factors, psychological factors were reported by some interviewees, which included anxiety from peer pressure, negative emotions caused by task sequence and distractions from testing settings. The interviewees' perceptions of task complexity is illustrated in Table 6.4.

Table 6.4 *Students' Perceptions of Task Complexity*

Category	**Perceptions of TC**	**Explanations**	**Interviewees**
Overall measurement	Task 3 is the easiest in the 4 tasks	Familiarity with the tasks	All
	Task 6 was easier than Task 4 and Task 5	Clear structure	In 5
		Familiarity with the task	In 2, In 3, In 6
	Task 6 was the most complex	Lack of prior knowledge	In 1
		Mental fatigue caused by task sequence	In 4
	Task 4 was the most complex	Extra information load	In 5 & In 8
	Task 5 was the most complex	Psychological pressure	In 3
		Lack of familiarity	In 6
Steps involved	Less complex	The reading step as a prior knowledge provider	In 1 & In 2 In 3, In 4
		Similarity between three-steps and daily classroom practice	In 2
	More complex	Extra information load	In 5 & In 8
Prior knowledge	Campus life situations	Association with their university life	In 1, In 2, In 3, In 5, In 6, In 7, In 8
	Academic lectures	Universal accepted without cultural difference	In 4
Task types	Stating- opinion tasks are the easiest	Clearly conveyed ideas	In 5
	Presenting solutions are complex	A lot of information load	In 2 & In 3
		Clear structure	In 1
	Presenting solutions are easy	The Problem and the solutions are given	In 8
	Illustrating a concept was the most complex	Extra information load	In 1, In 3, In 5
		Unfamiliarity	In 7
Planning time	Longer planning time reduces TC	Detailed planning and in-depth preparations	In 2, In 3, In 5, In 4 In 7, In 8
	Shorter planning time reduces TC	Peer pressure	In 4
		distraction	In 6
	Planning time is affected by task variability	Familiarity	In 3 & In 4

Note. TC = Task complexity; In = Interviewee.

6.2.2.1 Overall judgment

The interviewees sequenced the overall task complexity of the four tasks as: Task 5 > Task 4 > Task 6 > Task 3. They judged Task 3 as the easiest, because they felt familiar with the task which involved a conversation about university shuttle policies close to their university life. Similarly, as most of the interviewees were majoring in finance and international trade and were familiar with the topic on money in Task 6, they believed that Task 6 was easier in comparison with Task 4 and Task 5. For instance, Interviewee 6, like Interviewee 2 and Interviewee 3, connected his rating of Task 6 as an easier task to his familiarity with the task topic:

> I rated Task 6 as easy, because I think the words, sentences in the lecture are very familiar to me, and they match my major. You know, I come from the school of Finance and International Trade, and I know some information on money just like this task. (Interviewee 6/28/06/2018)

From Interviewee 6's quotes, it is not hard to find out that his prior knowledge of the task made the terms and jargons about money easier to be understood. His professional knowledge created his familiarity and reduced task complexity. Interviewee 7, the high-proficiency learner also emphasized the importance of familiarity:

> My understanding of today's tasks is that even you have good language proficiency, if you have no background knowledge, you will find the tasks are very difficult, because you don't' know how to develop your ideas. However, if you are very familiar with the topics, you will know how to express your ideas. Familiarity determines task complexity. (Interviewee 7/02/07/2018)

The importance Interviewee 7 attached to prior knowledge suggested that without relevant prior knowledge of a given task, even high-proficiency language learners would perform badly as they may not know how to respond to tasks only with their language skills. The importance of prior knowledge was further evidenced by the case of Interviewee 1 who, unlike most of the interviewees, rated Task 6 as the most difficult one. He recounted that he was not able to effectively and immediately process the information in the listening materials, because Task 6 did not provide reading materials

which he regarded as background information for familiarising the lecture. Apart from familiarity generated by prior knowledge, psychological factors also affected interviewees' rating. One example came from Interviewee 4 who blamed task sequence for his mental fatigue:

> I think Task 6 is the most difficult one, because it is the last task. Then I felt tired, thinking, wow, it almost finish! I didn't want to continue anymore! It's enough! So, I think Task 6 is the most difficult because when the task was given, I felt too tired to perform it. (Interviewee 4/27/06/2018)

Apparently, in the case of Interviewee 4, Task 6 was rated the most difficult not because of the task per se but because of its sequence. As the final task, Task 6 seemed to exhaust the interviewee mentally due to his intensive concentration on prior tasks. Another example came from Interviewee 3, one of the two interviewees who commented that Task 5 was the most difficult. The interviewee talked about her psychological pressure in task performance:

> Task 5 was completely different from what I predicted before I started the test. It was so difficult for me, and I had kind of disappointment, which caused pressure on me. (Interviewee 3/24/06/2018)

In performing Task 5, the interviewee had a positive expectation of the task at the beginning, which indicated her use of planning. However, when she found that the task was not what she had predicted, she felt disappointed, which caused pressure over her. Consequently, she perceived the task as complex.

6.2.2.2 Planning time

With regard to planning time, five respondents supported longer planning time, believing that it would lead to better performance, two interviewees were in favour of shorter time due to psychological factors, and one interviewee thought that the length of the planning time would be affected by task complexity. Interviewee 2 was one of the long-planning-time supporters. She held the view that the longer the planning time is,

the more that task performers could think about their responses in detail such as about the choice of vocabulary:

> I think 30 seconds allows me to make fuller preparations for the task. For instance, what I can do in 20 seconds is to prepare an outline. But if I am given 30 seconds, I can have a more careful thinking about the minor details of my responses. I can even think about the vocabularies for better answering the questions. So, the longer planning time I am given, the more possibly I can perform the tasks better, because the longer planning time renders me extra time to reflect and to think in depth. (Interviewee 2/24/06/2018)

Interviewee 2's perception indicated that longer planning time such as 30 seconds against 20 seconds would create more thinking space for task-takers. In the additional 10 seconds, task-takers could do more preparations based on what they did in the 20 seconds. If they could finish outlining in 20 seconds, with the 10 additional seconds they could polish the outline, think about better vocabularies to speak, and what's more, they could think of every detail of their responses to the tasks. It was clear that longer planning time could allow task-takers to process more useful information for the speaking tasks. When the overall information load was reduced due to the longer planning time, task-takers like Interviewee 2 tended to judge tasks less challenging and they would feel more confident in performing tasks.

On the other hand, Interviewee 4 and Interviewee 6 spoke for shorter planning time. Interviewee 4 elaborated his opinion by saying that if the four speaking tasks were taken into account as a whole, he believed that 30 seconds for planning should be provided in Task 3 and Task 4, because longer planning time could help test-takers warm up. After the warm-up, he commented that, no matter how long the planning time would be, either 30 seconds or 20 seconds, it made no sense to test takers like him. Additionally, the interviewee put forward an unexpected point:

> Well, personally speaking, I would rather support 20-second planning time. You know, in real testing situations, it is very possible that I just use only 20 seconds to think about what I should do for the tasks given. For the remaining 10 seconds, I might think of nothing. So, the

> remaining 10 seconds is waste of time to me. To make things worse, while I am forced to waste the 10 seconds and do nothing, my peers and classmates may use the time to think actively, and to take notes. In such circumstances, I will be more nervous in addition to the nervousness caused by the tests. Consequently, I will speak illogically, and get bad performance. Nevertheless, if I am given 20 seconds for planning, I will feel at ease, because I know when I cannot plan well in the given time, my peers are almost staying in the same boat, as they won't have the additional 10 seconds to plan better than me. The sameness will give me a peaceful mind, which, compared with a nervous brain, will help me in task execution. (Interviewee 4/27/06/2018)

Interviewee 4's preference for shorter planning time was due to his anxiety from peer pressure. In his view, the longer planning time may provide more competitive edges to his peers if he cannot make full use of it. Shorter planning time, in contrast, will give him the advantage in competing with his peers, because he believes that shorter planning time would deprive them of the advantages in preparations, and put them in the same situation which he feel more confident about. In other words, the longer planning time "threatened" the interviewee because of his worries on the possibility of the more effective and efficient use of the planning time by his peers than him. Such worries reflected his nervousness about peer competition. If the thread from peer competition was absent, he may have a different standpoint. That explained why when being further asked him which he would select, 30 seconds or 20 seconds, if there was no peer competition, the interviewee responded, without hesitation, that it was no doubt that he would select the longer planning time:

> If there is no competition from my classmates, it is for sure the longer the planning time is, the more fluently I would speak, and the more attention I would pay to the language such as conjunctions to avoid the sudden elapse in my speech which may cause by the shorter planning time like the 20 seconds. 30-second planning time, I believe, will enable me to make better logic connection between sentences in my speech. (Interviewee 4/27/06/2018)

The interviewee's changing preference for longer planning time appeared a little individual-centred or even a little selfish. However, the change showed that it was the psychological elements replacing task factors that affected him: The subjective psychological elements mirrored by his anxiety seemed to have a dominant role over the objective task factors such as planning time, exerting significant influence on his perception of task complexity.

Unlike Interviewee 4 who expressed his flexible views due to internal psychological pressure, Interviewee 6 reported his worry on distraction brought about by both internal and external factors, which contributed to his support of shorter planning time:

> For me, I think 20–second planning time is better than 30-second planning time, because in 30 seconds, I was easier to be distracted and I could not concentrate on the tasks. My mind went drift in 30 seconds, and I did not think of anything that was useful to finish tasks. When planning time was longer, I felt hard to do tasks well. (Interviewee 6/28/06/2018)

For Interviewee 6, what he concerned was the possible distraction caused by longer planning time. He sounded unconfident in his capability of concentrating on a task for a longer period of preparing time. Such lack of confidence revealed his internal mental fatigue: His mind was likely to unconsciously "go drift" after 20 seconds. So, he did not support 30-second planning time.

In addition to the above two quite opposite views on planning time, a neutral standpoint was held by two interviewees who advocated that in judging planning time, tasks per se should be taken into consideration. For example, Interviewee 3 believed that the effectiveness of planning time relied on tasks. Taking Task 5 as an example, she pointed out that no matter how long the planning time was, it made no sense:

> The task was so difficult, and no matter how long planning time I was given, it was useless to me, because I was not able to comprehend what the two speakers were talking about at all. The conversation completely differed from my prediction, so I felt very disappointed. (Interviewee 3/24/06/2018)

The female student's experience and opinion suggested that when a task was too difficult for task-takers, the length of the planning time did not function as expected. When she had no idea about the tasks, it was impossible for her to make any effective planning regardless of the planning time, as she might not know how to start the tasks. Even though she started the task based on her prior knowledge, the interviewee was not sure whether she would perform the tasks smoothly. Under such conditions, it was unavoidable that she felt difficult or even in vain using the planning time effectively.

Despite her disappointment in performing Task 5, Interviewee 3 admitted that 30 seconds would be better than 20 seconds for task performers in making a general plan before answering questions. However, she emphasised again in the end that planning time could not determine a task-taker's final performance, instead, the familiarity with tasks matters:

> If a task-taker knows nothing about the tasks or he or she is unfamiliar with the tasks, planning time does not make any sense. Even though longer planning time is accessible, it may be useless. What is more important is not the length of planning time, but the task-taker's familiarity with the tasks. (Interviewee 3/24/06/2018)

Interviewee 3's comments showed that her judgment of planning time should be interpreted from two perspectives. In a broader situation, she believed that the longer planning time benefits task performers like her, since they can have more time to think and plan for a better performance. Nonetheless, in some exceptional situations where tasks are too complex, planning time plays a unhelpful role. In consequence, task-takers may think that planning time is helpless, and they will turn to other resources to perform tasks. Familiarity, in this sense, is considered as superior over planning time for task-takers to rely on just as Interviewee 3 advocated.

6.2.2.3 Prior knowledge

When asked what kinds of tasks were easier, tasks on campus life or on academic lectures, seven interviewees perceived the former ones as more familiar to them, because they were university students, and had some prior knowledge on campus situation. One of the interviewees explained:

> Tasks on campus life involve activities that we have experienced as university students. Even though we had not experienced the activities, we have kind of familiarity. In the contrary, tasks on academic lectures were about something we have no prior knowledge of. Also, the vocabularies in the lectures were not familiar to us and they caused problems in our understanding. So, we felt that lectures were abstract and hard to understand. It was very difficult to link academic lectures to what we had learned. (Interviewee 2/24/06/2018)

Obviously, to these student interviewees, tasks on campus life were closely related to their daily routine, so it was easy for them to develop familiarity with the tasks. This familiarity accordingly assisted them to overcome negative responses, which justified why they rated campus life tasks as easier ones. Similarly, as they had no prior knowledge of lectures, they perceived academic lectures as more abstract and ranked lectures above campus life conversations on the task complexity scale. However, Interviewee 4 thought that due to culture differences between China and the United States (the campus life presented in the test tasks was situated in American universities), it was not an easy job for a Chinese EFL learner like him to connect the campus life to what he was familiar with in Chinese universities. He commented that tasks on campus life were a little complicated to comprehend because of his lack of prior knowledge of American universities:

> I think tasks on campus life are more complex, because campus life in a foreign university is too far away from us. If, for example, the campus life is about religion, food or a unique culture, Chinese students like me may not understand. So, it is extremely hard for us to develop a kind of sympathy with foreign campus life. (Interviewee 4/27/06/2018)

By contrast, he thought that academic lectures were universal, hence were easier to be understood:

> …But there are not so many culture biases in academic lectures. En…. It may not be called bias…it should be better called cultural differences. As lectures are universal and they remain the same in all countries.

> Anyway, they are still about prior knowledge. (Interviewee 4/27/06/2018)

From Interviewee 4's perspective, academic lectures were about universal knowledge regardless of geographic differences, whereas campus life was about the culture across countries. So, tasks on campus life rendered Chinese EFL learners less prior knowledge than tasks on academic lectures, which increased task complexity. Irrespective of the disagreement between Interviewee 4 and the other interviewees, one thing is clear that all the university student interviewees believed that prior knowledge reduced task complexity.

6.2.2.4 Steps involved

Most interviewees believed that compared with task involving two steps (listening and speaking), tasks that require task-takers to take three steps (reading, listening and speaking) before finishing the task were easier to be performed. The reason was that they considered the reading passage as an effective step to familiarise them with the given tasks. Interviewee 3 and Interviewee 6 even understood the reading step as a pressure reducer, thinking that if they had been required to speak immediately after listening without reading, they would have known nothing about the tasks, which would lead to pressure. Interviewee 3 interpreted how she considered such pressure:

> When we were given a reading material, we would find tasks easier to understand compared with listening to something that we had no knowledge of, isn't it? Listening to something new without a reading material as background information, we would feel huge pressure, and particularly, if the listening materials are completely unfamiliar to us, we might feel more pressure, and we would feel harder to perform the tasks. (Interviewee 3/24/06/2018)

In addition to taking reading as a pressure reducer, Interviewee 3 shared another seemingly neutral view on the reading step. She pointed out that whether tasks with three steps were difficult or not compared with those with two steps was, to a large extent, determined by task complexity. She took Task 3 and Task 4 as examples to support her view:

> I don't think the reading step make sense in performing Task 3, because it is so easy. Honestly, the reading material offered in this task was useless to me. However, when I did Task 4, the reading section was useful, because it gave me a lot of information on the lecture, otherwise I knew nothing about it. (Interviewee 3/2/06/2018)

Interviewee 3's opinion conveyed a message that whether additional reading made tasks complex depended on if it could provide task-takers with the prior knowledge they needed. Like Interviewee 3, Interviewee 6 also recognised the significance of the reading step:

> When we had access to the reading materials related to the tasks, we associated what we were reading with what we had learned and made purposeful prediction on the tasks. Because of this, I firmly believe that reading reduced task complexity. (Interviewee 6/28/06/2018)

Interviewees who voiced the positive role of reading step also included Interviewee 2 who advocated that reading helped her understand tasks and make psychological preparations for the tasks. Moreover, this interviewee supported the idea shared by Interviewee 3 and Interviewee 6, that is, task complexity was not determined by the steps it involved as she argued that:

> For today's tasks, the three steps were presented in a natural and logic order, which helped me gradually engage in the tasks rather than leading me directly to the points as the tasks with two steps did. Tasks with two steps didn't give me enough time to fully prepare. Although I had to go through three steps and had to process increasing information load compared with the tasks only two steps, I believe tasks with three steps were easier. (Interviewee 3/24/06/2018)

6.2.2.5 Task type

A large proportion of the interviewees judged the task of stating opinions as the easiest, because they thought that ideas in these tasks were stated clearly, hence easier for them to comprehend. Interviewee 5's view represented such a judgment:

> Task 3 required us to state one of the speaker's opinions, and it was easier to perform, since the opinions were stated very clearly. En… What I meant was that the two opinions of the two speakers in the conversation were clearly conveyed for us to understand. (Interviewee 5/28/06/2018)

The clear ideas in the interviewee's understanding may reflect the clear structure of the task. This made it easier for the students to acquire the key points for the speaking tasks. Two interviewees held negative attitudes to the tasks on presenting solutions, commenting that they involved a lot of information load, increasing task complexity and imposing pressure on task-takers as Interviewee 2 reported:

> Tasks on presenting solutions were more difficult because it was impossible that a task-taker like me could provide a relevant solution within the short preparing time given." (Interviewee 2/24/06/2018)

Interviewee 3 agreed to Interviewee 2's comments by adding that Task 5 required task-takers not only to understand but also to solve the problems encountered. The increase of information load made Task 5 more complex. In contrary to such negative attitudes, two interviewees had positive opinion on Task 5, for they believed that task-takers could refer to their own personal experience to answer the task questions, which reduced tasks complexity. To explain this view, Interviewee 1 made a comparison between Task 5 and Task 4:

> My solution was based on my understanding of the problem in Task 5. Such a task was flexible, and I could use my prior experience. So, I only needed to process small amount of information load. Yet tasks like Task 4 that asked us to illustrate an academic concept were not flexible, as an academic concept usually represents a sort of academic authority. To perform such tasks, I couldn't use my personal ideas, so I lost the flexibility and had to process more information. (Interviewee 1/24/06/2018)

The above opinion suggested the flexibility in Interviewee 1's approach to dealing with tasks. He believed that he could use his prior knowledge to finish the tasks confidently. The interviewee also mentioned about the close link between information load and his

prior knowledge: Prior knowledge leads to less information load. Another interviewee, Interviewee 8 added his judgment on the easiness of the presenting-solution task as the following:

> The presenting-solution tasks did not require making up solutions. Instead, they provide us some solutions as well as the reasons. So, in performing these tasks, what we needed to do was only repeating the solutions. (Interviewee 8/15/08/2018)

Obviously, interviewee 8 contributed the easiness of the presenting-solution task to the fact that solutions and the reasons had been provided in the task. On the task about illustrating an academic concept with examples, half of the interviewees thought that the task was more difficult than the other tasks because of the unfamiliarity and the huge information load. Interviewee 7 was one of them:

> I believe that even an academia may find it hard to explain an academic concept, let alone task-takers like us who had no prior knowledge at all. (Interviewee 7/02/07/2018)

His remarks showed that tasks on illustrating an academic concept involved academic knowledge increased task complexity. In fact, Interviewee 7's case seemed to indicate that the source of task complexity may not be the task type, but the academic knowledge on the tasks. If task-takers do not know such prior knowledge, they might believe that tasks are complex as the above interviewees experienced. Despite their mixed views on task complexity due to varying task types, interviewees generally agreed upon the influence of prior knowledge on task-takers. Five interviewees reported that regardless of task type, familiarity with tasks played a critical role in their rating of task complexity as Interviewee 7 reported earlier. They argued that if a given task could fit task performers' prior knowledge, it would be easier for the task performers to develop associations between their prior knowledge and the task. As a result, they would find less difficult to perform the task and thereby they would have better performance. Interviewee 3's comments provided an additional example:

> My feeling on complexity related to task type was that task complexity depended on, to a large degree, our prediction prior to tasks, and our prior knowledge as well. Put it another way, whether our prior

> knowledge match a task closely related to our judgment of task complexity involved in that task. (Interviewee 3/24/06/2018)

Due to the salience of prior knowledge in her task execution, Interviewee 3 also pointed out her understanding of the correlation between prior knowledge and task type:

> I don't think task type makes sense in determining task complexity. Instead, the information that tasks can give task-takers makes sense. If the information is familiar to them, a task, regardless of its type, can be accepted by task-takers, otherwise, the task may be hard for task-takers to perform. (Interviewee 3/24/06/2018)

Obviously, Interviewee 3 weighed prior knowledge over task type, and in her view, discussion about task type makes no sense if prior knowledge is not taken into consideration in judging task complexity. Interviewee 4's view shares similarity with that of Interviewee 3, but his interpretation sounded more conclusive:

> No matter what task type we are supposed to face: Stating opinions, presenting solutions or illustrating a concept with examples, the key factor that decides task complexity is whether task performers have relevant prior knowledge on the tasks. Without it, we will have a lot of limits in terms of the ideas that we can use to perform the tasks. (Interviewee 4/27/06/2018)

The arguments of the two interviewees revealed that to these students, they attached more importance to prior knowledge than to task type. In their opinion, the presence of prior knowledge is likely to help them overcome the hurdles caused by various task types.

6.3 Teachers' Responses

Under the category of each individual task, eight themes were identified in the teachers' responses to the open-end questionnaire. Among them, six themes were about task complexity, and two were related to metacognitive strategy use. The eight themes included: (a) The presence or the absence of prior knowledge; (b) reasoning demand reflected by task type; (c) easily-understood task content; (d) clearly-cut task structure; (e) speech speed in the recordings; (f) increased information load; (g) planning; and (h)

monitoring. Table 6.5 displays the Chinese EFL teachers' responses and the relevant themes.

Table 6.5 *Responses and Themes related to Teachers' Perceptions of Task Complexity*

Respondents	**Perceptions**	**Task 3**	**Task 4**	**Task 5**	**Task 6**
T 1	Responses	Prior knowledge and task type reduce task complexity	Prior knowledge reduces the task complexity	No prior knowledge increases task complexity	Task type increases task complexity
	Themes	The presence of prior knowledge and Reasoning demand reflected task type	The presence of prior knowledge	The absence of prior knowledge	Reasoning demand reflected by task types
T 2	Responses	University buses are not popular in most Chinese Universities, but the content of the notice and the man's ideas are easily understood.	This is common topic most students feel confused and they would like to discuss. So, the preparation is not difficult.	The girl is facing a problem and the professor offers solutions. But it takes time to the girl's choices.	The professor gave two definitions of money, and the example she used is easy to understand.
	Themes	Easily understood task contents	easier preparations thanks to prior knowledge	Reasoning demand reflected by task type (planning)	Clearly-cut structure

Respondents	Perceptions	Task 3	Task 4	Task 5	Task 6
T 3	Responses	The words used are familiar to examinees on campus, so they have no problems giving a complete answer.	Spotting and understanding terminologies in an academic lecture make the speaking tasks more difficult. As long as examinees spot those terminologies and understand in the context, they could do the speaking task well. Otherwise they might miss some points or details. Note-taking and sentence organization is decisive. And examinees are supposed to speak more than in Task 3.	More information to be processed in the situation. Complexity of speech is obviously higher than the previous task.	More specific information provided and to be processed, which caused more mental work and language complexity in understanding of the task and answering the questions.
	Themes	The presence of knowledge	the absence of prior knowledge; note-taking (monitoring) and sentence organization (planning)are suggested.	Increased information load	Increased information load

Respondents	Perceptions	Task 3	Task 4	Task 5	Task 6
T 4	Responses	University bus service and student's driving are not so popular in China, but the content of the notice and the man's opinion are easily understood. The difficult point is to support the man's opinion by combing the two aspects: Bus route and parking space.	Even though the term of audience effects is not familiar to students, they must have had the similar experience mentioned in the material, so that it is not difficult for them to connect the two situations in the material with their own life experience.	The situation is not popular in most Chinese universities. The difficult points are as follows: Figuring out the Mary's situation and her decision of what to do next and why; the conversation is a little difficult for Chinese students to follow because of its topic and speed.	This type of English lecture is not very difficult for Chinese students to follow, even though sometimes they are subject-related. Most Chinese students are familiar with the content mentioned in this lecture.
	Themes	Easily understood task contents	The presence of prior knowledge	The absence of prior knowledge, speed	The presence of prior knowledge

Note. T = Teacher interviewee.

6.3.1 Task complexity

As shown in Table 6.5, prior knowledge was the most frequently explained factor that influenced the teachers' judgment of task complexity. For instance, without prior knowledge on economy reflected in Task 6, the teachers regarded it more complex than Task 4 which, they believed, offered prior knowledge related to their daily experience. The availability of prior knowledge was the explanation of the higher rating of Task 6 than Task 4 given by the teachers.

Task type reflecting reasoning demand ranked the second as the source of task complexity. Teacher 2 supported such ranking, who advocated that unfamiliar task types reflected a task's reasoning demand, and they required more time on task-takers to prepare or to plan. In fact, such unfamiliarity with task type also indicated the absence of prior knowledge. In commenting the source of task complexity, Teacher 2 and Teacher 4 believed that Task 3 was easy because of its easily understood contents. Teacher 3 attributed task complexity involved in Task 5 and Task 6 to the increased information load. "clearly-cut structure" and "speech speed" were used by Teacher 2 and Teacher 4 respectively as the sources of task complexity of Task 6 and Task 5. With reference to the criteria, although the two teachers gave same ratings to Task 4 and Task 6, the overall means of the teachers' ratings (see Table 5.11) evidenced more complexity in Task 6 than in Task 4. The task sequence in terms of task complexity perceived by the teachers' is: Task 5 > Task 6 > Task 4 > Task 3. Such sequence was not consistent with that perceived by the student interviewees but was in conformity with that judged by the students engaged in the quantitative phase.

6.3.2 Metacognitive strategies

Although the teachers were originally invited only for their judgment of task complexity, two themes on metacognitive planning and monitoring emerged from their responses. These themes additionally revealed the close association between task complexity and metacognitive strategy use. The teachers believed that when prior knowledge, the critical factor determining task complexity which was also recognised by the interviewees, was not accessible, task complexity might increase. In the circumstances, task-takers might refer to metacognitive strategies for meeting the challenge. Teacher 3's suggestion on the use of note-taking, one form of monitoring

strategy, and organising, one form of planning strategy, to deal with Task 4 exemplified their view on the association.

In summary, although the teachers' perceptions of task complexity in the four tasks differed from what the student interviewees as stated earlier, agreement was reached in the two groups: Prior knowledge, compared with the other three task factors under investigation, exerted more effects in determining task complexity. Furthermore, the teachers proposed that the increase of task complexity generated by the unavailability of prior knowledge could be dealt with via appropriate use of metacognitive strategies.

Chapter Seven
General Discussion

7.1 Chapter Overview

Research findings in line with relevant theoretical issues are discussed in this chapter, through which the Chinese EFL learners' online strategic competence and performance in response to task complexity in the context of the computer-mediated integrated speaking test are better understood. The chapter starts with the discussion on the Chinese EFL students' strategic competence typically represented in their metacognitive strategy use across tasks, and their perceptions and the Chinese EFL teachers' evaluations of task complexity. Following this, the complex relationships among the three research variables are discussed.

7.2 Individual Metacognitive Strategy Use across Tasks

Given the students' rather high level of language proficiency as stated in Chapter 4, the integration of the data from the questionnaire and the semi-structured interview suggests that the students did not use the individual metacognitive strategies actively across the four integrated speaking test tasks, and their use of the strategies was affected by task complexity and their individual attributes. Considering the complexity of the individual metacognitive strategy use across the test tasks identified in the study, discussion along this direction is proceeded from two perspectives: A holistic view of the four individual strategy use and an analytic view of the specific individual metacognitive strategy use.

7.2.1 A holistic view

As noted above, the students are not active users of the individual metacognitive strategies when their language proficiency and the descriptive analysis of their individual metacognitive strategy use are considered. The result is obviously unexpected, because the positive correlation between language proficiency and strategy use has been evidenced in a large volume of studies on L2 learning strategy use (Oxford, 2004; Psaltou-Joycey & Gavriilidou, 2018), especially in those that focus on university students (e.g., Griffiths 2013, Vrettou 2009, 2011) as was the case in this study.

The inconsistency between the present study and previous literature may be due to the students' lack of training on metacognitive strategies. It is known that if individuals are able to use strategies, they need to learn them just as they must expose themselves to language in order to acquire it (Fazilatfar, 2010). Daily classroom instructions on the selection and the appropriate use of metacognitive strategies in the target language setting have been proved as an effective means to enhance EFL learners' awareness of metacognitive strategy use, and to empower them to use metacognitive strategies appropriately in light of the given context (e.g., Oxford, 2017; Qin, 2018; Takeuchi, 2020; Zhang & Zhang, 2019). Grounded in this understanding, if the students in this study had the learning experience in metacognitive strategy use, it would be very likely that they would have resorted to metacognitive strategies in performing the speaking test tasks.

Unfortunately, both the students and the teachers reported that they had no access to metacognitive activities in their classroom instructions, and they had no relevant knowledge of metacognitive strategies at all. Such a report is not surprising since it is common that in the Chinese educational context, EFL teachers are not asked to teach their students how to use different strategies, particularly, the metacognitive strategies as stated in Chapter One. Hence, EFL learners and even some EFL teachers have no knowledge of metacognitive strategies and of the appropriate use of these strategies when dealing with language tasks (Dabarera, Renandya, & Zhang, 2014; Oxford, 2017). Just as a Chinese idiom says, "巧妇难为无米之炊" (qiǎo fù nánwéi wúmǐzhīchuī, in English, it is impossible to make bricks without straw), it was impossible for the students to actively use metacognitive strategies when the repertoire of the strategies was not readily available due to a lack of learning. Noteworthy is an empirical study (Pei, 2014), which revealed that metacognitive strategy instructions did not have a significant impact on Chinese EFL learners' metacognitive strategy use in reading. Nevertheless, the result may have to do with the reading domain in which the study was conducted (de Boer et al., 2018; Oxford, 2004), and hence it cannot disvalue the salient role of the metacognitive strategy instructions in empowering L2 learners to use metacognitive strategies actively and effectively.

Another possible explanation for the finding in this subsection relates to the students' perceptions of task difficulty across the four TOEFL-based integrated speaking test tasks. Descriptive analysis indicated that all the four tasks were judged by the students as very difficult (see Table 5.10). The judgment, to a large degree, is consistent with Yu et al.'s (2017) finding that Chinese EFL learners perceived TOEFL integrated speaking test tasks as difficult. Because of the judgement, the students might have realised that their strategic efforts would not help them complete the test tasks. Consequently, they might not bother to think out and use appropriate strategies to take the test when they sat the test voluntarily without any specific goals as some interviewees reported (Oxford, 2017).

What is more, the students' appraisal of task difficulty may weaken their confidence in completing the speaking test, causing their anxiety and loss in motivation **(**see Section 6.2.1). In consequence, their interest in engaging in the test was likely to decrease (Li, Lee & Solmon, 2007). In learning activities, learners engage in tasks emotionally and cognitively. Emotional engagement refers to the learners' affective reactions to learning activities and learning environment, including the level of their interest, sense of boredom and anxiety; and cognitive engagement is the approaches taken by the learners to tasks, which highlights the learners' efforts to employ cognitive learning strategies to regulate the learning process. The two types of engagement work interactively (Lynch, et al., 2019). In line with the literature on engagement, when the students lost engagement in the test emotionally (e.g., interest, confidence, motivation and anxiety) due to task difficulty, their cognitive engagement was likely to be negatively impacted, which accordingly affect their strategy use including metacognitive strategy use. Therefore, it is understandable why some interviewees reported that they did not set any goals, the use of the metacognitive planning, because the test tasks were too difficult (see Section 6.2.1.1).

In fact, the inactiveness in metacognitive strategy use across tasks demonstrated by the high-proficiency students has long been documented in empirical studies on the correlations between L2 language learner's language proficiency and their communicative strategy use (e.g., Chen, 1990; Ellis, 1984a, 1984b), among which Chen (1990) is the most similar to this study because of the participants' backgrounds. Through a qualitative interview data analysis on 12 Chinese EFL speakers in two groups

(high-level proficiency group versus low-level proficiency group), the researcher investigated the participants' communicative strategy use in relation to their language proficiency and identified the negative correlation between the two variables. In addition to strategy use, Chen attributed the finding to the wider resources and the more linguistic knowledge accessible to the high-proficiency Chinese EFL speakers compared with their low-proficiency counterparts. The findings lend some support to this study. Similarly, Oxford (2004) analysed that owing to L2 participants' high language proficiency, their strategy use became an unconscious process that was habitual and automatic, and therefore might not be reported on a strategy use survey. It is apparent that Oxford's analysis also provides some supporting evidence for this study.

7.2.2 An analytic view

Regarding each specific metacognitive strategy use (planning, problem-solving, monitoring and evaluating), descriptive statistics suggests that problem-solving was reported as the most frequently used strategy while monitoring was the least reported strategy. The result is somewhat surprising in that problem-solving is not the subcomponents of the more widely acknowledged and applied model of metacognitive strategies comprised of planning, monitoring and evaluating (see Section 3.2.5).

7.2.2.1 Problem-solving

Students reported that they used "inference" and "substitution" (the two subcomponents of the problem-solving strategy defined in this study as shown by Table 3.1) the most frequently. Such a result may have to do with the way in which the students performed the integrated speaking testing tasks. According to O'Mally and Chamot (1990), EFL learners tend to use strategies in a problem-solving manner, so it is possible that the students considered their use of various strategies as application of the problem-solving strategy and reported them on the inventory. Indeed, in line with some scholars (e.g., Flavell, 1979; Goh, 2008; Papaleontiou-Louca, 2008; Tarricone, 2011; Zhang, 2017), the students' understanding of using the problem-solving strategy reflects their metacognitive knowledge of strategies. As EFL learners' metacognitive knowledge may be fallible or false (Brown, 1987; Zhang, 1999), it is therefore likely that EFL learners believed that they used the problem-solving strategy in performing the four tasks given and report it in the inventory. However, as their knowledge of the strategies might be false, their reported use of the problem-strategy might be the problem-solving strategy

and hence the high frequency in their use of this strategy reflected in the inventory might not be true (Brown, 1987; Veenman, Hout-Wolters & Afflerbach, 2006). The fallibility related to the Chinese EFL learners has been reported by Zhang (1999) whose study disclosed the fallibility of Chinese university EFL learners' metacognitive knowledge associated with their reading strategies.

On the other hand, data from the interviews indicate that the students' preference for using the problem-solving strategy could be linked to their learning experience. According to the interviewees, in their EFL learning activities, they were often requested to practise the use of substitution and inferencing by their English teachers. Such classroom instructions are common in China where the pedagogical focus of EFL teaching is on students' language proficiency as stated in Chapter One. In such situations, educated guesses or inferences are one of the basic strategies which Chinese EFL teachers emphasise in teaching reading and listening, while substitution is the strategy that Chinese EFL learners are taught to deal with the speaking and writing tasks, although these teachers may not be aware that they are teaching metacognitive strategies. As a result, inferencing and substitution are treated as two of the compulsory strategies that Chinese EFL learners are required to master in their daily EFL learning activities (Sun, 2016; Zhou, 2020). The pedagogical actuality has permitted the student participants to immerse in a problem-solving strategy learning environment, which made it natural that the students used the strategy frequently in performing the test tasks. Chinese EFL learners' preference for using inferencing and substitution as evidenced in this study was also reported in Yang and Sun (2012), and Yin (2013). In these studies, when investigating the use of substitution and inferencing, among other metacognitive strategies in writing, the researchers discovered that the two problem-solving strategies worked more effectively in Chinese EFL learners' performing speaking tasks than writing tasks. This phenomenon is indeed not unique to Chinese EFL learners, as, according to Halliday and Matthiessen (2013), EFL learners generally tend to use substitution in performing speaking tasks. The tendency may explain why substitution and inferencing, against the other three metacognitive strategies, demonstrated the highest frequency in the students' strategy use to process the speaking tasks. The relationship between the students' preference in metacognitive strategy use and speaking also confirms findings reported in Cohen (2018) who summarised that L2 learners' strategy use could be categorised in line with a specific language skill or

modality, which put strategy use in a well-placed position. Additionally, such a relationship coincides with the view held by Oxford (2004, 2017) who commented that the use of L2 learning strategies is associated with a specific language skill area.

7.2.2.2 Planning

Of the three subcomponents designed to denote the students' use of planning on the questionnaire, setting goals affected the students' planning use the most compared with self-management and organizational planning. According to the interviewees, overall, it was because of their motivation and test anxiety that they did not plan, and hence set no specific goals.

Motivation is a complex and multidimensional goal-driven activity. It is one of the most important individual factors that affect L2 learners' strategy use (Cohen, 2018; Oxford, 2017), which has been empirically supported by researchers (e.g., Berger & Karabenick, 2011; Martínez, Pérez & Navarrete, 2015). In the well-accepted dichotomy of motivation in L2 learning proposed by Deci and Ryan (2000), intrinsic motivation refers to learners' behaviours that can bring them gratification without thinking about the consequence of their behaviours such as learning a language for the joy of learning per se. In contrast, extrinsic motivation, another constituent of the dichotomy, functions like a stimulus in learners' learning process through which learners can receive external rewards such as getting an ideal job or being admitted to a university. In this study, since the students were volunteers, it is obvious that they did not sit the integrated speaking test for selecting a course or for a job. In such situations, it is justifiable to think that the students were generally unmotivated extrinsically. On the other hand, their voluntary participation seemed to indicate that they had intrinsic motivation. Despite this, the actuality was that except for Interviewee 1 and Interviewee 3 who reported their participation as a chance of oral practice which demonstrated their intrinsic motivation, overall, the interviewees expressed no motivation in sitting the speaking test. The obvious conflict between the participants' voluntary participation that suggested their intrinsic motivation and their self-report of no motivation through the interviews leads to one question: Why were the students not motivated? Some interviewees attributed this to task difficulty, complaining that the test tasks were too difficult, hindering their interest to perform the test. The students' association of task difficulty with motivation points to a concept: Task motivation.

Task motivation was originally proposed by Julkunen (1989) who defined the concept as a composite of trait motivation and state motivation. Trait motivation refers to L2 learners' general motivation as discussed above, while state motivation means how the learners perform in a task (Kormos & Wilby, 2019; Ma, 2009; Poupore, 2013; 2015; 2016). According to Kormos and Wilby (2019), L2 learners' task motivation is influenced by their self-efficacy, expectancy-value, intrinsic motivation and interest. Self-efficacy is the learner's belief that they are competent enough to complete a given task or accomplish a specific goal. It is a critical determinant of the learners' motivation. Expectancy-value regards the learners' value of tasks and their expectations of successful task completion. Intrinsic motivation and interest are interchangeable terms (Ajzen, 2002; Horvath, Herleman & Lee McKie, 2006), and both are linked to the learners' internal feelings in performing tasks.

Many researchers hold the view that task motivation elements are immediately connected to learners' perceptions of task difficulty: The increase of task difficulty perceived by learners typically lead to the decrease of their self-efficacy, expectancy-value, intrinsic motivation and interest (e.g., Dörnyei, 2014; Dörnyei, Kormos & Wilby, 2019; Li et al., 2007; Ma, 2009; Poupore, 2016). In line with this view, the students might experience very weak self-efficacy, expectancy-value, intrinsic motivation and interest, and hence generated low even no task motivation in performing the test tasks, as they perceived the four integrated speaking test tasks as extremely difficult. Since task motivation is the macro aspect of learners' motivation (Dörnyei, Kormos & Wilby, 2019; Poupore, 2016), given the possible effect of motivation on strategy use and goal setting as discussed previously, it is reasonable that the students did not set any specific goals for these speaking test tasks. This result can borrow some support from the study of Poupore (2015, 2016) who investigated learners' motivation in interactive tasks and found that task complexity and individuals' familiarity with the task contents influenced the individuals' motivation. In a similar study by Masrom, Alwi and Daud (2015), negative correlation between motivation and task complexity was found in 88 high-proficiency Malaysian university students on the computer-mediated communication writing tasks. Considering the similarities in both the participants and the research contexts, Masrom et al.'s empirical study offers additional support for the finding of this subsection. The effect of motivation on Chinese EFL learners' metacognitive strategy use is also documented elsewhere in the literature (e.g., Bai & Guo, 2019; Chang & Liu,

2013; Xu, 2011). For instance, Bai and Guo (2019) discovered relatively high correlations between participants' motivation and their strategy use from the data collected on 523 primary school students in Hong Kong. As for the complicate relationship in task difficulty, motivation and strategy use, through a study on 163 university students in different majors in Taiwan by means of questionnaires and reading and listening tests, Chang and Liu (2013) concluded that Chinese EFL learners who perceived the task given as less difficult demonstrated strong motivation and higher frequency of metacognitive strategy use. This result obviously lends convincing support to this study. Additionally, the close association between task engagement and task motivation discussed earlier (see Section 7.2.1.1) also accounts for the possibility of the unfavourable effect of the students' perceptions of task difficulty on their motivation in taking the test tasks.

Another factor that may have a negative relationship with the students' goal-setting is test anxiety which was unexpectedly reported by the interviewees who had relatively higher language proficiency. Though conflicting with some literature that suggests inverse variability of test anxiety across test-takers' language proficiency (e.g., Lien, 2016; Liu, 2018), this finding is similar to Aydin (2013) and Karatas, Alci, and Aydin (2013) who found that Turkish EFL learner's proficiency level was not the determinant of their anxiety level. Likewise, Huang and colleagues' (2013) study showed no significant correlations between test-takers' language proficiency and their test anxiety in the integrated speaking test tasks. The researchers attributed the result to the voluntary participation of the test-takers. However, in the present study, the test anxiety reported by the higher-proficiency students may come from the evaluation pressure or their fear of poor performance deriving from their performance expectancy in the speaking test (Saha, 2014; Zeinder, 2010). Performance expectancy refers to test-takers' expectancies of failures due to bad test performance (Zeinder, 2010). According to Zeinder (2010), in performing test tasks, compared with those with low language proficiency, higher-proficiency students tended to have higher expectancy of their own performance. However, when they found that the integrated speaking test tasks were difficult and hence beyond their competence, it was easier for them to worry about the possible unexpected poor performance. To avoid such expectancy failure, they might not set any goals to reduce their expectation of themselves and consequently to reduce the test anxiety from which they might suffer.

7.2.2.3 Monitoring

Monitoring was reported as the least frequently used metacognitive strategy, and this finding concurs with Swain et al. (2009), Barkaoui et al. (2013), Huang (2013), and Yi (2012), the four studies that bear the closest relevance to this research. In these studies, metacognitive monitoring was either not used at all or used the least frequently by participants. Yet, this strategy was found to be frequently used in similar studies on reading (e.g., Phakiti, 2016; Purpura, 2013; Zhang & Zhang, 2013), listening (e.g., Goh, 2002; Nett et al., 2012; Rukthongand & Brunfaut, 2020;Vandergrift & Goh, 2012), and writing (e.g., Karlen & Compagnoni, 2017; Wischgoll, 2016). A possible explanation of this mismatch is the complex speaking process. As reviewed in Section 3.4.3, monitoring is one of the four key stages in a speaking process and it engages in the whole process of speaking in either a covert form or an overt manner (Bygate, 2011; Kormos, 2006, 2011). Since the students had no prior knowledge of metacognitive strategy use, it is possible that they were not aware of their actual use of monitoring when the strategy functioned in their speaking process in an overt form. Consequently, when they were reporting their metacognitive strategy use in processing the speaking test tasks on the questionnaire, they could not truly retrospect their use of monitoring. In addition, Barkaoui et al. (2013) have argued that compared with other language skills or modalities such as reading and writing, speaking has special requirements on speakers because of the immediate and online characteristics of speaking performance. Henceforth, in the integrated speaking test, the students had to process a rather huge load of information that involved not only speaking but also reading and listening. In addition, they also had to deal with the time constraint and pressure, and the immediate and online requirements on performance. The challenges from the speaking test provided few chances for the students to take useful notes, consciously apply learned or self-developed rules, or relate information to personal experiences when the students had no knowledge of metacognitive strategy use and of the speaking test tasks. Put another way, their unfamiliarity with the test tasks and the test task demands made it hard for the students to monitor their speaking process in a covert and effective manner.

Despite the low frequency of monitoring reported by the students on the questionnaire, the responses from the interviewees suggested a positive correlation between monitoring and L2 test-takers' language proficiency: Those who reported the frequent use of monitoring were the students with higher language proficiency. This positive correlation

has been evidenced in ample literature (e.g., Oxford, 2017; Phakiti, 2016; Teng & Zhang, 2019; Zhang et al., 2019). Roebers, Krebs and Roderer (2014), for instance, concluded that high-proficiency EFL learners were more accurate in using monitoring after a study on children's metacognitive process in relation to their language proficiency in a testing context via structural equation modelling. Alternatively, Ekhlas and Shangarffam (2013) examined Iranian university EFL learners' use of monitoring in taking the IELTS and the positive association between the participants' metacognitive strategy use and their language proficiency was identified.

7.2.2.4 Evaluating

Evaluating was reported by the students as the less frequently used metacognitive strategy in comparison with problem-solving and planning. This result is consistent with the findings by Zhang, Goh and Kunnan (2014) and Saraswati (2017) on L2 reading, of Nett et al. (2012) on multiple subjects including English, and of Rukthong and Brunfaut (2020) on L2 integrated listening. Some researchers have argued that the low frequency is related to the time constraint that the participants had to challenge in testing or to the participants' low language proficiency (e.g., Saraswati, 2017), while others considered the complexity of the metacognitive strategy use and the participants' lack of training as possible explanations (e.g., Nett et al., 2012). In this study, apart from time constraint and the students' lack of relevant training discussed previously, possible explanations are the students' confounding of evaluation with their metacognitive experience, the number of items regarding evaluating on the questionnaire, and the interdependent correlations existing in the four metacognitive strategies.

As data analysis of the interviews revealed, the students' self-evaluation was mainly built upon their subjective feelings: Their feelings of familiarity with the test tasks, their feelings of task complexity, their feelings of satisfaction with and confidence in their performance. According to Efklides (2002, 2006, 2008), those subjective feelings, were essentially the students' metacognitive feelings, one components of metacognitive experience. Due to this, in reporting their use of evaluating, as the items on metacognitive evaluating in the metacognitive strategy inventory were developed and validated in light of the definitions of metacognitive evaluating rather than the metacognitive experience, it is very likely the participants were not able to find the items that match their understanding of evaluating strategy reflected by metacognitive

experience. As a result, the items on the inventory that were used to elicit true metacognitive evaluating strategy might not be ticked by the participants, and accordingly the use of evaluating use might demonstrated low frequency.

In fact, in the five items related to evaluating on the questionnaire (see Table 5.7), only two of them are about post-task evaluation on the students' performance that might be used to elicit their subjective feelings. As for the other three items, they are related to checking goal setting and examining errors in test performance. Because the students did not set any goals, it is unsurprising that when recalling their use of evaluating, they were not able to report any activities elicited by the two items on goal-setting, and almost all the interviewees placed much emphasis on the subjective feelings of their performance. Similarly, due to the time constraint and the high demand of the speaking performance in assessment, it is unlikely that the students would make efforts in examining the possible errors in their performance, as test-takers typically had to rush at the end of a test (Pan & In'nami, 2015). As a result, they might not report error examination on the one item in the questionnaire. Taken together, it is reasonable that the descriptive analysis of the monitoring use resulted in low frequency count.

Moreover, in accordance with Efklides (2006, 2008), individuals' metacognitive feelings are drawn from their monitoring on the features of the task-processing experienced and/or the outcome of the processing. This opinion indicates that individuals' metacognitive feelings are affected by their use of monitoring, which further elucidates the close relationship between monitoring and evaluating. The relationship between the two metacognitive strategies has been recognised and statistically evidenced by Purpura (1999). Likewise, in O'Mally and Chamot's (1990) model of metacognition (see Section 3.2.2), self-monitoring was also closely associated with self-evaluation. Based on this relationship, since monitoring was reported by the students as the least frequently used strategy, it is tenable that the students experienced similar low frequency in their use of evaluating.

7.2.3 Metacognitive strategy use, individual attributes, and task complexity

In essence, the above discussion illustrates that the students' metacognitive strategy use was influenced by their individual attributes (e.g., motivation, anxiety, and learning experiences). This finding is supported by wider literature on L2 learning strategies. For

example, O'Mally and Chamot (1990), and Fazilatfar (2010) propose that many factors affect an individual's use of strategies including goals, prior learning experiences, task demands or task complexity and personal motivation (Dörnyei,2005, 2014, 2019). In L2 assessment, the direct effects of individual attributes on strategy use have also long been acknowledged (e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; Fulcher, 2015a, 2015b; Kyle, 2020; O'Sullivan, 2000; 2012, 2017; Weir, 2005). Bachman and Palmer (1996, 2010), for instance, believe that individuals' attributes such as motivation and anxiety may combine with test tasks, influencing their use of strategic competence, either facilitating or limiting the flexibility with which test-takers responds in a specific context. In the same vein, O'Sullivan (2011, 2017) postulates that a task-taker's psychological characteristics (e.g., motivation and emotional states) and personal experience (e.g., education and examination experiences) interact with test tasks, affecting test performance. From the perspective of task complexity, the relationship between individual difference and task complexity has been recognised by Gilabert (2000, 2007), Robinson (2007), Robinson and Gilabert (2020), Skehan (2018) and Révész et al. (2016).

However, the relationship between metacognitive strategy use and the students' individual attributes is partially contradictory with Bachman and Palmer's (2010) framework of the non-reciprocal language use where test-takers' individual attributes are regarded as the peripheral characteristics of test-takers in contrast with their strategic competence which is given a core role in the framework (see Section 2.2.2.3). Given the fact that the framework and the strategic competence model are still in the process of validation which needs further empirical supporting evidence (see Section 2.3.2.1), these effects are justifiable, but they point out the critical role of the students' individual attributes in L2 speaking assessment.

7.3 Task Complexity across Tasks

Discussion about task complexity across tasks is conducted from two levels: The synergetic effects of the task complexity factors and the effects of the individual task complexity factor on the overall complexity across tasks.

7.3.1 Synergetic effects

The substantial variance in task complexity perceived by the students across tasks shown in the one-way repeated measures ANOVA suggests the variability in the synergetic effects of the three task complexity factors (planning time, steps involved and prior knowledge) along the resource-dispersing dimension across the four tasks. This result lends empirical evidence to Robinson's (2015) Triadic Componential Framework in which when variables along the two dimensions of task complexity are manipulated simultaneously, the overall complexity of a specific task is assumed to vary, indicating the synergetic effects of such simultaneous manipulation. The result also conforms to some studies on the effects of simultaneously manipulating task complexity factors on task complexity and task performance (e.g., Gilabert, 2004, 2007a, 2007b; Levkina, 2008; Levkina & Gilabert, 2012). In a like manner, the variability in the overall task difficulty across task types perceived by the students supports Foster and Skehan (1996) who advocated that different task types impose various amount of cognitive load on task performers, generating varying degrees of task difficulty. The students' perceptions of task type also coincide with Gan (2012), Madarsara and Rahimy (2015), and Rezazadeh et al. (2011) who postulate that variability of task types leads to variability of task complexity.

Data analysis on the one-way repeated measures of ANOVA also showed that the students rated Task 5 as the most complex task, followed by Task 6, Task 4 and Task 3. The rating was in agreement with that of the Chinese EFL teachers reflected in the qualitative analysis of data collected on the open questionnaire. The agreement confirms Révész et al.'s (2016) study where teacher experts' judgments on task complexity demonstrated a highly positive correlation with EFL learners' ratings. The agreement indicates the validity of the participants' perceptions of task difficulty. The sequence of the four tasks is also consistent with Robinson's (2015) Triadic Componential Framework. When a task (e.g. Task 5) provides less planning time and involves various task types, it will impose more difficulty on task performers; while a task (e.g. Task 3) which gives prior knowledge, more planning time and involves fewer task types will be easier for task performers to complete.

However, the sequence was not consistent with the judgments of the student interviewees who rated Task 5 as the most complex, followed by Task 4, Task 6, and

Task 3. It is noticeable that the inconsistency lies in the participants' different perceptions of task complexity involved in Task 6 and in Task 4. Data analysis in ANOVA and the open questionnaire supported more cognitive complexity in Task 6, whilst data from the semi-structured interviews indicated its less complexity. This confrontation may be explained by the sampling in the study: Most of the student interviewees were majoring in finance and international trade. Their prior knowledge of finance possibly elucidates why they perceived Task 6 as easier in comparison with Task 4 although both tasks were about academic lectures. Task 6 was on an academic concept of money related to finance, about which most of the interviewees had prior knowledge, whereas Task 4 was about a psychological term which seemed unfamiliar to the interviewees. In essence, the inconsistency in the participants' perceptions of task complexity suggested that the availability of prior knowledge might reduce task complexity, which corroborates with much of the literature (e.g., Ellis et al, 2019; Robinson, 2015, 2018; Robinson & Gilabert, 2020; Skehan, 2018).

7.3.2 Individual factors

The above disagreement on task sequence also revealed the participants' perceptions of the role of prior knowledge in the four test tasks: Both the students and the teachers agreed that in all the four task complexity factors, prior knowledge was the determinant of a task's complexity. They further claimed that if prior knowledge of a given task was available, task performers would have familiarity with the task, and under such circumstances, they could make effective use of the planning time to make a prediction and make a guess even when they were not able to completely comprehend the task input. On the other hand, if task performers had no relevant prior knowledge, they were likely to believe that the task was difficult to perform, and in such situations, planning time made no sense and even a waste of time, which might negatively influence task performers (e.g., test anxiety and worry about distraction) as recalled by some interviewees. This view is in light of Huang et al.'s (2016) and Huang et al.'s (2019) studies in which test-takers experienced test anxiety when they performed an integrated speaking test task on an unfamiliar topic. The participants' view on prior knowledge also echoes Wang and Yu (2018) who found that Chinese university students automatically used their background knowledge in understanding the tasks given and whether the topics of the tasks were related to the participants' prior knowledge would affect their performance.

When assessing task complexity established by task type, the students gave their lowest rating to Task 3 which required test-takers to narrate the speakers' opinions, followed by Task 6 and Task 4, the two tasks to do with justifying an academic concept with two examples in a lecture. By contrast, they gave the highest rating to Task 5 which involved decision-making and justification. The results show that the narrative task was perceived by the students as easier compared with the justification task and the decision-making task. The rating partially vindicates the study conducted by Foster and Skehan (1996) who investigated task complexity in three types of tasks (personal information exchange, narrative and decision-making) under the task conditions of +/-planning time and their impacts on learners' oral performance. Their study disclosed that narrative tasks were less cognitively demanding compared with decision-making tasks. Furthermore, the results borrow some support from Rezazadeh, Tavakoli, and Rasekh (2011) who asserted that descriptive (narrative) tasks are easier than argumentation (justification) tasks in writing.

With regard to the effect of the steps involved in the four integrated speaking test tasks on task complexity, data analysis in ANOVA showed that the two tasks that required more steps (reading, listening and speaking) were measured by the students as the most difficult (Task 5) and the easiest (Task 3). With reference to the task characteristics of the four speaking test tasks (see Table 3. 2), both Task 5 and Task 3 were about campus life, indicating the availability of prior knowledge to the students, and hence the difference between the two tasks lay in task type and planning time. Since compared with Task 5, Task 3 provided longer planning time (30s versus 20s) and simpler task type (narration versus problem solving plus justification), it is highly likely that Task 5 was perceived as more difficult than Task 3, although both tasks involved three steps. The result apparently supports Robinson's (2015) framework. Additionally, the students rated Task 5 as more difficult than Task 4 and Task 6, which also conforms the framework. On the other hand, they rated Task 3 as easier than Task 4 and Task 6, and this rating seemed to conflict with the framework. In addition to the intricateness of the simultaneous manipulation of multiple task complexity factors, explanation of the conflict could be that the interviewees considered the additional step (reading) in the speaking test tasks were less challenging because they treated the step as prior knowledge provider. As reported earlier, since either the students or the teachers prioritised prior knowledge as the most important determinant of task complexity in the

four task factors under examination, it is expected that the two three-step tasks were evaluated as less difficult than they were assumed to be within Robinson's framework.

Collectively, the participants' perceptions of planning time, prior knowledge, steps involved only partially agreed to Robinson's (2011a, 2011b, 2015, 2018) hypothesis on task complexity: The independent presence of planning time and prior knowledge leads to less complexity in a task; and more steps in comparison with less steps indicates almost the similar cognitive load on the students. The two groups of participants' perceptions of task complexity also reveal that when the synergetic effects of the variables are taken into account, the presence or absence of prior knowledge determines, the participants' measurement of task complexity caused by the variance of the other three task complexity factors. This result is borne out well with the proposal of Oxford et al. (2004) who posited that familiarity with tasks makes a big difference for a learner in considering whether a certain task is to be complex. Newton and Nation (2020) also supported this view on the importance of prior knowledge in determining task complexity. As reviewed previously (see Figure 3.3), such a role of prior knowledge, among the four task factors, in determining task complexity suggests its moderating impact on the relationships between the other three task variables and task complexity. Figure 7.1 illustrates the moderator role of prior knowledge.

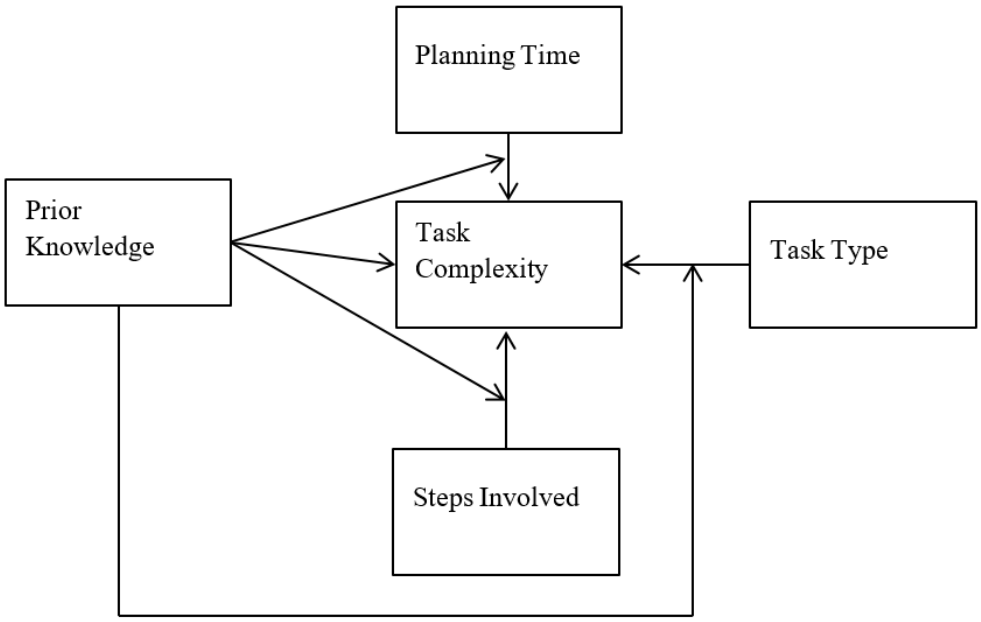

Figure 7.1 *Prior Knowledge as a Moderator*

7.4 Metacognitive Strategy Use and Task Complexity

In terms of the relationships between individual task complexity factors and individual metacognitive strategy use, descriptive statistics of the questionnaire showed that the students used more problem-solving and planning on the cognitively easier tasks than on the cognitively complex tasks due to the availability of prior knowledge and planning time. The finding is consistent with the study by Pan and In'nami (2015) which revealed that test-takers tended to use planning frequently when given cognitively easier tasks. It is also partially corroborated by Barkaoui et al. (2013), Swain et al. (2009) and Yi (2012) who reported that Chinese EFL learners frequently used planning when tasks provided them with long planning time and more prior knowledge. The role of prior knowledge was further illustrated by the students' more frequent use of monitoring on tasks with unfamiliar topics, which supported the teachers' responses to the open questionnaire: When prior knowledge is inaccessible, Chinese EFL learners may use monitoring to perform tasks. Regarding metacognitive strategy use across task types, the students stated that Task 5, though rated as the most complex task, allowed them to associate task demands with their personal experience. Relating information to one's experience is one of the subcomponents of monitoring under investigation, which explains the rather high frequency of monitoring use on Task 5. With respect to evaluating, the students reported higher frequency use of the strategy on easier tasks involving two steps than on more complex tasks involving three steps. This finding is consistent with that of Barkaoui et al. (2013), Swain et al. (2009) and Yi (2012) which identified the more frequent use of evaluating on listening-speaking tasks than on reading-listening-speaking test tasks.

In general, the slight statistical variance in the students' use of the individual metacognitive strategies across tasks illustrates the effect of task complexity on test-takers' use of individual metacognitive strategies, though not substantial. The result coincides with the finding by Barkaoui et al. (2013), Swain et al. (2009), Yi (2012) and Oxford and her colleagues (2004), in which the types and frequencies of the strategies used by participants were not found to be significantly affected by task complexity. Despite such coincidence, the result contradicts Ikeda and Takeuchi's (2001, 2003) study where a significant effect of task complexity on participants' reported use of reading strategies was identified. Possible explanations of the contradiction could be the

differences in student samples, the strategies and the tasks between the Japanese researchers' study and this research. In Ikeda and Takeuchi's study, participants are Japanese EFL learners, and the tasks used to trigger the strategies are reading texts. By contrast, the student participants in this study are Chinese EFL learners and the tasks employed are the integrated speaking test tasks. Furthermore, the strategies under investigation in this study are the metacognitive strategies in the testing context as opposed to the metacognitive reading strategies in the participants' daily language learning. Simply put, the variability in research methodology leads to the contradictory research findings (Corriero, 2017; Creswell & Creswell 2018).

Regarding the synergetic effects of the task complexity factors on metacognitive strategy use, the output of the one-way repeated MANOVA, and the high correlation index between metacognitive components suggest that the interactive metacognitive strategies reported by the students demonstrated substantial variability in response to the changing task complexity. The result implies that metacognitive strategies operated in a clustering manner and they were task dependent, which supports the existing literature. For instance, Nett et al. (2012) whose longitudinal study with an experience sampling analysis on 70 Germany students showed that the metacognitive strategies used by the participants worked interactively in test performance. In addition, Fernandez (2018) unfolded the employment of the clusters of the metacognitive strategies reported in IELTS speaking tests via coding participant's discourse. The finding also concurs with Rukthong and Brunfaut' s (2020) study where metacognitive strategies were used concurrently by participants in their listening task performance. Finally, the actual working mode of the metacognitive strategies also lends some support for the more recent study conducted by Wang and Yu (2018), in which Chinese EFL learners' cognitive activities underwent iterative performance in testing conditions.

7.5 Metacognitive Strategy Use and Speaking Performance

Empirical studies on the relationships between metacognitive strategy use and test performance (see Section 3.5.1) have been extensive yet inconclusive (e.g, Aryadoust & Zhang, 2016; Fernandez, 2018; Nourdad & Ajideh, 2019; Pan & In'nami, 2015; Phakiti, 2018; Rukthongand & Brunfaut, 2020; Sawaswati, 2017). This possibly accounts for why the current study resulted in weak correlations between the students'

metacognitive strategy use (either in a clustering manner or in an individual form) and their test performance. The finding reflects many researchers' views on Bachman and Palmer's (2010) model of strategic competence: The model needs additional empirical validation (e.g., Phakiti, 2018; Song, 2005; Zhang & Zhang, 2013).

A possible alternative explanation is the instrument employed to elicit the students' metacognitive strategies. Some researchers have pointed out that although self-report questionnaires have witnessed an extensive application in measuring metacognitive strategies, they may not represent what the participants actually do (e.g., Greene & Azevedo, 2010; Jacobse & Harskamp, 2012; McNamara 2011; Schellings & Van Hout-Wolters, 2011; Veenman 2011). In addition, the weak correlations may be caused by task complexity. When the students found the speaking test tasks were extremely difficult, they might turn to whatever resources available to deal with the testing tasks. Under such conditions, it is common that the students used more strategies on all the four difficult tasks, and consequently, their test scores could not display considerable changes across the four tasks as reported by Barkaoui et al. (2013) and Swain et al. (2009). Moreover, since L2 speaking assessment is a complex process, the finding can lend some credibility to Bachman and Palmer's (2010) proposal that in addition to metacognitive strategy use, many factors affect test scores such as individual attributes, test methods and random measurement errors. Indeed, as discussed previously, the fact that the students did not experience considerable changes in their strategy use across the four tasks may also account for this result.

In the current literature on L2 assessment, the weak relationship between metacognitive strategy use and test performance has been discovered in many other similar studies. For example, in examining the relations between individual metacognitive strategy use and test-takers' integrated speaking test performance, Barkaoui et al. (2013) and Swain et al. (2009) found no significant and positive relations between the two variables. Similarly, Fernandez's (2018) study showed no positive correlation between strategy use and participants' test performance reflected by their test response quality in the IELTS speaking test tasks. In terms of other language skills, Phakiti (2003, 2018) and Purpura (1997, 1998) revealed either weak or no connections between EFL learners' use of metacognitive strategies and their reading test performance. Pan and In'nami (2015) also reported a weak relationship between the two variables: Metacognitive strategy

uses only accounted for 7% of the variance in test scores in participants' listening performance. Additionally, Song's (2005) study that involved reading, writing and listening skills led to a similar finding.

Irrespective of the weak association, problem-solving was found to significantly affect test performance on Task 5, the most difficult task. This result may be due to task complexity rather than strategy use. As discussed earlier, the students preferred to use more strategies when encountering difficult tasks, and they used problem-solving the most frequently. Following this, it is understandable that the frequency of problem-solving use was reported as the highest on Task 5. However, this positive relationship conflicts with the study by Yang and Sun (2012), where the use of the problem- solving (e.g., substitution) did not correlated noticeably positively with Chinese EFL learners' performance in writing tests. A possible reason for such opposite results is the variability in research methodology: The different instruments (speaking tests versus writing tests) through which the correlation between problem-solving and test performance was investigated (Corriero, 2017; Creswell & Guetterman, 2019). The variance in the students' strategy use caused by the difference in testing instrument regarding language modalities was also supported by Cohen (2018) and Oxford (2004, 2017), as discussed in Section 7.2.1.1.

7.6 Task complexity and Test Performance

As task complexity was statistically indicated by the students' self-rated task difficulty, discussion about the effects of task complexity on test performance mainly focuses on the synergetic effects of the four task complexity factors on the students' test scores.

Generally, the strong negative relationship between task complexity and the students' oral scores aligns well with Bachman and Palmer's (2010) framework of non-reciprocal language use in terms of the effects of test task characteristics on test performance. The negative relationship also corresponds to Robinson' (2015) Triadic Componential Framework regarding the effects of task complexity on task performance. Furthermore, the relationship validates Kormos' (2011) Bilingual Speech Production Model where tasks, as input, determines speech production.

Empirically, the finding has been confirmed by an impressive body of literature on L2 assessment (e.g., Huang, 2013; Rahimi & Zhang, 2018, 2019; Robinson, 2007; Sasayama, 2016). For instance, in researching the effects of listening test tasks on test-takers' performance, Pan and In'nami (2015) found that four different test tasks with varying degrees of complexity triggered different scores in participants: When tasks became more difficult, the participants' test scores decreased accordingly. Similarly, Zhang et al. (2014) discovered the variance in Chinese EFL learners' test performance when they performed four reading tasks characterised by different difficulty. Additionally, the findings of the three studies that bear the closest relevance to the current research (see Section 3.5.4) showed that Task 4 and Task 5 elicited the lower test scores from the test-takers than Task 6 and Task 3 (Barkaoui et al., 2013; Swain et al., 2009; Yi, 2012). Such findings support this study in which the students were granted the lower test scores on the more difficult tasks (Task 5 and Task 6) than on the easier ones (Task 3 and Task 4). In the research field of tasks, the influence of task complexity on task performance has been researched extensively with almost universal results: Negative correlation between task complexity and task performance exists. More specifically, the studies on the synergetic effects of the simultaneous manipulation of multiple task complexity factors on performance, though just a few, lead to such a correlation (see Section 3.5.2), lending support for this study.

7.7 Metacognitive Strategy use, Task Complexity and Speaking Performance

The moderating effect of the individual monitoring strategy on the relationship between task difficulty and oral scores discovered in HLM conflicted with the proposal on the mediating role of metacognitive strategies in L2 assessment (e.g., Bachman, 1990; Bachman and Palmer, 2010; Barkaoui et al., 2013; Chapelle, 1998; Swain et al., 2009). The conflict may be attributed to the validity of Bachman and Palmer's (2010) strategic competence model upon which the proposal is built (Barkaoui et al., 2013; Chapelle, 1998; Swain et al., 2009). As discussed previously, the mixed results generated from the exiting literature has challenged the validity of the model, which warrants further validation (e.g., Phakiti, 2003; 2008; 2016; Purpura, 1997, 1998, 2008; Seong, 2014; Zhang, 2017; Zhang et al., 2014). Now that the strategic competence model is still in the process of validation, the moderating role of the metacognitive strategies is tenable.

However, the moderating effect of the students' strategic competence in performing the integrated speaking test is well supported by Kormos' (2011) Bilingual Speech Production Model in which L2 speakers' strategic competence or their metacognitive strategy use moderates the correlation between speaking tasks and speaking performance (see Section 3.4.3). The finding is further corroborated by Liu and Li (2011, 2012) who pointed out that the complexity-performance relationship is moderated by many factors including task-takers' cognitive abilities such as their strategy use. Empirically, this finding conforms to the results of several studies on the similar research topic (e.g., Barkaoui, 2013; Huang et al., 2018; Hung, 2013; Swain et al., 2009).

While working as an important moderator between task complexity and test performance (Hayes, 2018), monitoring was found to be used the least frequently on all the four integrated speaking test tasks in the four metacognitive strategies as discussed previously. Possible explanations of such seemingly contradictory findings are: (a) The critical role of monitoring as a subcomponent of metacognition in language learning; (b) the importance of the appropriateness over frequency count in language strategy use; and (c) the indispensability of monitoring in L2 speech production.

In the research arena of metacognition, it is generally accepted that individuals' metacognition functions at their meta-level, and it relates to the objective world though monitoring and control (Efklides, 2001, 2008). In all the metacognitive strategies, monitoring has been reported as a strong predictor of an individual's academic performance (Shih & Huang, 2020). Such a view elucidates why Flavell (1979) labelled his metacognition model as the model of cognitive monitoring. In the L2 learning domain, monitoring operates in an omnipresent form, and is regarded as a key factor in an individual's learning process, as the strategy is essential in helping learners develop and understand complicated information in the learning process (Oxford, 2017; Qin, 2018; Zhang & Zhang, 2019). It is likely that due to such importance, monitoring, not the other three strategies which had higher frequency reported by the students, exerted a moderate effect on the association between task complexity and their test performance, weakening the negative influence of task complexity on their test scores.

The conflicting results on monitoring also reflect some researchers' standpoints on L2 learning strategy use such as Cohen (2014), Huang (2013), Oxford (2004, 2017), Plonsky (2019), Shih and Huang (2018) who advocated that appropriate use, rather than the frequency of strategies in response to a given task, should be considered as an indicator of successful strategy use. Alternatively stated, "measuring strategy use in terms of frequency count only furnishes part of the picture, and serous considerations needs to be given to the appropriateness of the strategy use for the given context" (Takeuchi, 2020, p. 75). With their empirical studies in L2 assessment, Huang (2013) and Rukthong and Brunfaut (2020) echo this opinion, pointing out that in L2 assessment, the effectiveness and the success of a test-takers' use of strategies is not determined by the frequencies of the reported strategies, instead, the appropriateness of the strategy use makes sense. This view lends support to the students' reported monitoring use in this study.

Furthermore, the result borrows some support from the literature on the omnipresent role of monitoring in L2 speech production proposed by Kormos (2006, 2011), Bygate (2011), and Luoma (2004). These researchers believe that monitoring is activated in the whole process of the speaking covertly and overtly, which might result in the unawareness of its existence by the students who had no prior training, knowledge or experience related to how to use strategies. Accordingly, the low frequency of the strategy use was reported as discussed previously. Because of this, the reported use of monitoring may not be the actual use of the strategy, which attributes to the conflict between the lowest frequency of monitoring use across tasks and its salient moderating effect on the task-performance relationship.

In terms of the clustering metacognitive strategies, they also had significant moderating impacts on the interaction between task complexity and the students' oral scores. However, the impacts were very likely to be generated by the individual monitoring. This is because when the substantive effects of the interactive metacognitive strategies on the interaction between task complexity and the students' scores were identified, monitoring constantly played an active and substantial role in the interactive metacognitive strategies in the HLM. The actuality suggests that the seemingly significant effects exerted by the interactive metacognitive strategies on the students' test performance may essentially relate to the moderating impact from the individual

monitoring strategy (Barkaoui, 2010, 2013; Garson, 2013; Wen, 2009; Raudenbush & Bryk, 2002).

A very surprising phenomenon should be highlighted. As one form of individual attributes, the students' metacognitive strategy use was not found to significantly affect test scores in the HLM. In spite of this, the results of HLM indicate that a half of the variance in the students' test scores at Level-1 could be explained by the variability in their individual attributes at Level-2. Such results support the qualitative findings on the essential role of test takers' individual attributes (not their strategic competence) in L2 assessment as discussed earlier, though the construct is beyond the research focus of this study.

Chapter Eight
Conclusion and Future Steps

8.1 Chapter Overview

This chapter begins with a summary of the research findings followed by a brief conclusion. It then presents the contributions and implications of the current study to L2 assessment and L2 education. Limitations which indicate future steps in relevant research and my final thoughts as the closing remarks are provided in the end.

8.2 Summary of Findings

With reference to the research questions (see Section 1.6.3), quantitative results showed that the metacognitive strategies reported by the Chinese EFL learners varied significantly in a clustering manner across tasks. Task complexity, on the other hand, could be indicated statistically by task difficulty, and it had significant effects on speaking performance. These results addressed Research Question 2, Research Question 4, Research Question 6 and Research Question 7. Regarding the relationships among metacognitive strategy use, task complexity and speaking performance, monitoring functioned as a moderator in the association between task complexity and speaking performance, which answered Research Question 8.

Surprisingly, among the four individual metacognitive strategies under investigation, variance was only found in problem-solving across tasks, which helped to address Research Question 1. Similarly, the four interactive metacognitive strategies had no substantial influences on speaking performance. Despite this, the independent problem-solving strategy had a significant effect on speaking performance on Task 5. The results answered Research Question 5. Moreover, qualitative findings revealed that the Chinese EFL learners were not active users of metacognitive strategies, but they used problem-solving actively; the learners' individual attributes, though not the focus of this study, served as a mediator in the correlations between metacognitive strategy use and task complexity. The findings supported and complemented the quantitative results concerning Research Question 1. In terms of task complexity, the qualitative findings showed that prior knowledge worked as a moderator between task complexity and the

other three independent task complexity factors. The findings answered Research Question 3. Figure 8.1 illustrates the relationships identified among the Chinese EFL learners' strategic competence in the form of their metacognitive strategy use, task complexity and the learners' speaking performance. In the figure, to distinguish the roles of monitoring and problem-solving, different colours were used. Additionally, the moderating role of prior knowledge is not shown as it has been displayed separately in Figure 7.1.

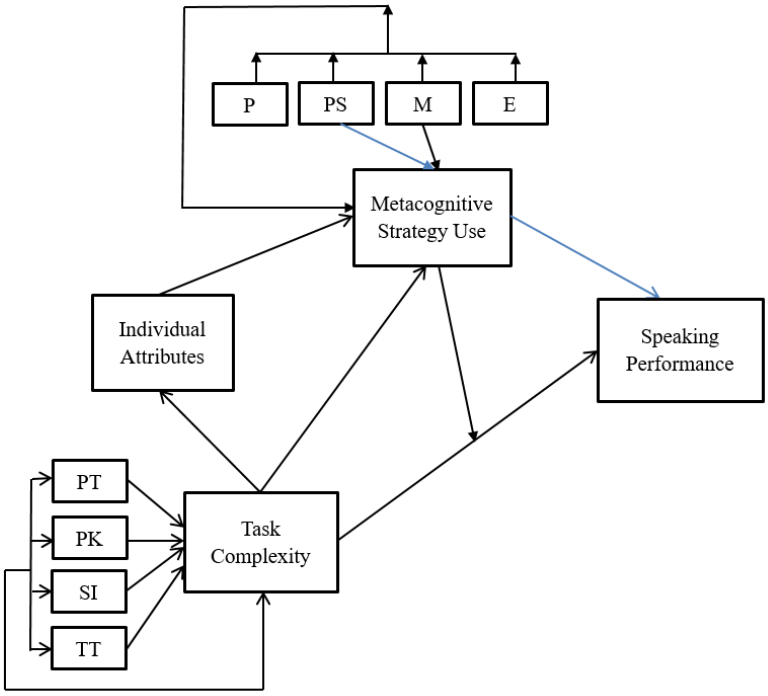

Figure 8.1 *The Identified Relationships in the Research Variables*

Note. P = planning; PS = problem solving; M = monitoring; E = evaluating;
PT = planning time; PK = prior knowledge; SI = steps involved; TT = task type.

8.3 Conclusion

This study was set up to examine the complex relationships among test-takers, test tasks and test performance within Bachman and Palmer's (2010) framework of non-reciprocal language use. A convergent mixed-methods research design in an exploratory approach from an interdisciplinary perspective was adopted for the examination. The research findings fit partially Bachman and Palmer's (2010) framework of non-reciprocal language use and lend some support for Robinson's (2015) Triadic Componential

Framework on task complexity and Kormos' (2011) Bilingual Speech Production Model.

More specifically, the study reveals that without relevant prior knowledge, EFL learners' strategic competence cannot respond appropriately to tasks with varying degrees of complexity in L2 speaking assessment, hence exerting no substantial effects on task performance as assumed by some researchers (e.g., Bachman & Palmer, 2010; Cohen, 2018, Oxford, 2017; O'Sullivan, 2013; Weir, 2005). By contrast, test task characteristics that reflect the cognitive complexity of the four speaking test tasks can elicit variability in EFL learners' individual attributes (e.g., their motivation and anxiety) which further stimulate various responses from their use of individual metacognitive strategies. In the meanwhile, test task characteristics have direct and significant impacts on both EFL learners' use of the interactive metacognitive strategies and their test performance. Nevertheless, the strong and negative influence of test tasks on test performance is moderated by monitoring, indicating that the use of metacognitive strategies, though not directly affecting test performance, helps EFL learners reduce the negative impact from task complexity on their performance in L2 speaking assessment.

The complex relationships among the three research variables investigated within the research context of Chinese EFL learners may present some insights on L2 speaking assessment from a Chinese perspective. Furthermore, the study is expected to fill the research gaps discussed in 1.3.2 and provide some new ideas in relation to metacognitive instructions and task-related implications in L2 testing and L2 learning as delineated in the subsequent subsections. In the meanwhile, it answered the questions from my classroom instructions: Both task complexity of the integrated speaking test tasks and my students' lack of strategy use learning contributed to their weak performance on the TOEFL-based integrated speaking test, though they had a rather high level of English language proficiency.

8.4 Contributions

While functioning as empirical evidence for validating Bachman and Palmer's (2010) framework of non-strategic competence model, Robinson's (2015) Triadic Componential Framework, and Kormos' (2011) Bilingual Speech Production Model,

the findings of this study potentially contribute theoretically and methodologically to the research practice in L2 assessment and L2 learning. To be specific, they will: a) Provide new insights related to Bachman and Palmer's framework of non-reciprocal language use; and b) present a valid and reliable research instrument for exploring metacognitive behaviours.

8.4.1 Contribution to theory

Theoretically, the research findings on metacognitive strategy use in relation to task complexity will provide some new insights into Bachman and Palmer's (2010) framework of non-reciprocal langue use in the following aspects.

8.4.1.1 Inclusion of problem-solving in the framework

Problem-solving, though typically not a subcomponent of metacognitive strategies in the models related to the metacognitive construct, was identified as a salient strategy in helping the students perform the test tasks. Such identification validates some researchers' (e.g., Chamot, 2005, 2009) views on the inclusion of the strategy in the metacognition model (see Section 3.2.5). It accordingly proves Bachman and Palmer's (1996, 2010) strategic competence model in their framework of non-reciprocal language use should be reconsidered and the problem-solving strategy is expected to be included in the model. The inclusion will serve as a research effort to respond to the proposal from some scholars (Aryadoust & Zhang, 2016; Phakiti, 2003, 2008, 2016; Purpura, 1997, 1998; Seong, 2014; Zhang, 2017; Zhang et al., 2014) in L2 assessment: Metacognitive strategies validated by empirical studies should be included in Bachman and Palmer's (2010) strategic competence model for its comprehensive validity.

8.4.1.2 Contextualisation of strategic competence in L2 assessment

With regard to the role of strategic competence in L2 assessment, it has been revealed by quite a few studies that mixed results were found in the effects of test-takers' strategic competence on their performance in the four modalities of reading, listening, writing and speaking (see Section 3.5.1). In this study, although the moderating effect of monitoring on the interaction between task complexity and speaking performance is consistent with Kormos' (2010) Bilingual Speech Production Model, it challenges the role of test-takers' strategic competence as a mediator in L2 assessment advocated by researchers including Bachman and Palmer (2010). Such different roles of strategic

competence as a mediator versus as a moderator indicate that test-takers' strategic competence is proposed to be examined in light with a specific language modality rather than with the general context of L2 assessment. As discussed previously that strategy use varies in response to different modalities (e.g., Cohen, 2018; Oxford, 2004), test-takers' strategic competence illustrated in Bachman and Palmer's (2010) framework may need to be contextualised in line with a specific language modality.

8.4.1.3 Equal importance of individual attributes to language ability in the framework

As the core component of language ability, strategic competence was affected, to varying degrees, by the students' individual attributes in response to task complexity. As the mediator between task complexity reflected in test task characteristics and strategic competence, individual attributes had correlations with both of the two key variables in L2 assessment. Hence, it is unconvincing to claim that language ability is the core part of personal characteristics weighing over other individual attributes (e.g., motivation, anxiety and metacognitive experiences reported by the student participants) as did by Bachman and Palmer (1996, 2010). From this perspective, the findings of this study show that individual attributes and language ability merits equal importance in Bachman and Palmer's (2010) framework of non-reciprocal language use.

8.4.2 Contribution to research methodology

As a widely recognised assessment tool of eliciting and examining individuals' metacognitive strategy use, questionnaires enjoy considerable popularity in L2 assessment and L2 learning. In striking contrast, only a few questionnaires on individuals' metacognitive strategy use are available. In addition, these questionnaires are either too task-specific, focusing on a specific language skill, such as Phakiti's (2003, 2008, 2016) Cognitive and Metacognitive Questionnaire on reading, Zhang and Goh's (2006) Metacognitive Awareness Inventory in Listening and Speaking Strategies on listening and speaking, and Vandergrift et al.'s (2006) Metacognitive Awareness Listening Quetionniare on listening, or too general like Oxford's (1990) Strategy Inventory for Language Learning related to non-testing daily learning (see Section 3.2.6). Due to this, decontextualised use of the above instruments is common, which is criticised by many researchers (e.g., Oxford, 2017; Taguchi, 2020). Such criticism obviously highlights the contribution of the newly developed Metacognitive Strategy

Inventory (MSI) to the research practice concerning the examination of strategic competence in L2 speaking assessment.

Moreover, quite a few L2 tests today are delivered by computer for efficiency such as the TOEFL speaking tests, one of the instruments deployed in the current study, when computer-mediated communication has become pervasive in L2 learning (Abdolrezapour, 2019; Chapelle & Voss, 2016; Farr & Murray, 2016; Isbell & Winke, 2019; Yu & Zhang, 2017). Against this backdrop, the availability of a valid and reliable questionnaire to investigate test-takers' strategic process in the context of computer-based testing such as the MSI is essential in that it can help test developers and language educators better understand test-takers' or learners' internal response to a given test task. Furthermore, although the context in which the MSI was employed is computer-mediated integrated speaking tests, such contextualisation does not exclude its possibility of being widely applied to explore speaking tests in any forms and the speaking activities in non-testing contexts due to the diverse sources from which the questionnaire was developed (see Section 3.2.6)

8.5 Implications for Practice

The importance of problem-solving and monitoring in L2 speaking assessment, and the substantial effect of task complexity on test performance are likely to provide pedagogical implications in Chinese EFL teachers' metacognitive scaffolding, syllabus designing and task development for classroom instructions on speaking. By the same token, the findings are likely to offer empirical validation evidence for task measurement and development in L2 speaking assessment.

8.5.1 Pedagogical implications

As stated above, implications for pedagogical purpose are related to metacognitive instructions, and syllabus designing and task development.

8.5.1.1 Metacognitive instructions

The mastery of metacognitive strategies has always been considered as crucial in one's sustainable learning (Oxford, 2017; Teng & Zhang, 2016; Zhang et al., 2016). Despite this, the acquisition of the metacognitive skills remains challenging for a learner if

external support or scaffolding (e.g., a teachers' help) is not available, because learning process is typically long lasting and non-liner. Consequently, metacognitive scaffolding, the most used scaffolding type in education, plays an essential role in learning (e.g., Dabarera et al., 2014; Zhang & Zhang, 2019; Zhang & Zhang, 2018; Zhang et al., 2019). In China, due to the actuality of EFL teaching where strategic competence has long been neglected (see Section 1.4), metacognitive scaffolding becomes more essential in classroom instructions (Finkbeiner et al., 2012; Jumaat & Tasir, 2016).

Metacognitive scaffolding, derived from metacognition, refers to instructional guidance that facilitates learners' metacognitive thinking to support their planning, monitoring and evaluating in learning (An & Cao, 2014). The underlying goal of metacognitive scaffolding is to assist learners in solving problems (Jumaat & Tasir, 2016; Xie & Bradshaw, 2008), and its effectiveness has been extensively acknowledged by researchers (e.g., Oxford, 2017; Qin & Zhang, 2019; Rahimi & Zhang, 2019; Teng & Zhang, 2019). Due to this, metacognitive scaffolding is proposed to empower learners to increase their self-efficacy for approaching learner autonomy, and to decrease their reliance on external assistance (An & Cao, 2014; Finkbeiner et al., 2012).

Amongst various means though which metacognitive scaffolding can be achieved, classroom instructions are asserted to be the most direct and most popular in the actual practice (An & Cao, 2014; Dabareraet al., 2014; Finkbeiner et al., 2012; Jumaat & Tasir, 2016). However, research efforts in metacognitive instructions have generated diverse and inconclusive views (Oxford, 2017; Plonsky, 2011, 2019), just as Oxford (2017) commented that "no strategy instruction guidelines were consistently applied and that the strategy instruction took place in many different situations and conditions and with disparate learners." (p. 312).

In the current study, problem-solving and monitoring strategy were identified as the essential metacognitive strategies used by the Chinese EFL learners in performing speaking tasks. It also evidenced that metacognitive strategies worked in a cluster. Built upon these findings, two types of metacognitive instructions are proposed in EFL classrooms, particularly in Chinese EFL speaking instructions. In this sense, the study is expected to provide additional insights on the development of a consistent strategy instruction guideline as commented by Oxford (2017) above.

The first type of metacognitive instruction is that EFL teachers may consider paying special attention to problem-solving and monitoring in designing the syllabus for classroom activities on speaking. By doing so, EFL teachers can familiarise their students with the use of the two strategies, allowing them to repeatedly practice such strategies. As a result, the EFL students' awareness of adopting the two metacognitive strategies in task performance, the preparatory but fundamental step for strategy learning, is likely to be raised (Rubin et al., 2007; Oxford, 2011, 2017). Also, since the two metacognitive strategies are proved to be effective in accordance with the Chinese EFL learners' retrospection, such metacognitive instructions are consistent with Oxford's (2017) proposal which underscores EFL teachers' attention to their students' cognitive needs based on students' feedback on strategy use. Furthermore, Plonsky's (2019) meta-analysis on strategy instruction also supports this type of metacognitive instructions, as his analysis reveals that when the target metacognitive strategies are narrowed down in classroom instructions, students tend to learn these strategies in the most effective manner. In China, teaching strategic competence to EFL learners in accordance with their preference for strategy use has also been advocated by scholars (e.g., Wang, Lai & Leslie, 2015).

The second type of metacognitive instructions relates to multiple-strategy instructions. Although metacognitive scaffolding can be realised through either single-strategy instructions (teaching students how to use a single strategy for aiding their learning) or multiple-strategy instructions (teaching students how to choose strategies from a set of strategies for effective learning) (Taylor, 2011), the latter form is preferred by many scholars (e.g., Akkakoson, 2013; Almasi & Fullerton, 2012; Dabarera et al., 2014). This is because multiple-strategy instructions enable learners to acquire a set of strategies to meet their actual needs in learning. Such instructions are particularly effective in helping learners develop their metacognitive awareness (Akkakoson, 2013; Dabarera et al., 2014; Pressley & Allington, 2014). The clustering characteristic of metacognitive strategy use identified in this study indeed coincides with the multiple strategy instructions. The coincidence implies that compared with the single strategy instructions, multiple strategy instructions might be prioritised in EFL teachers' classroom instruction for speaking so that their students can be more competent in performing speaking tasks before they approach self-regulation, the ultimate goal of education (Dabarera et al., 2014; Teng & Zhang, 2019). Yet, it has to be pointed out that it is more learner-friendly

and theoretically feasible for EFL teachers to apply this multiple strategy instructions after their students have acquired individual strategies as a foundation step (Chamot & Harris, 2019; Oxford, 2017).

To summarise, the two types of metacognitive instruction practice are shown in Figure 8. 2. In the figure, the blue arrows indicate the two metacognitive strategies to which EFL teachers should pay special attention in their single-strategy instructions as discussed above.

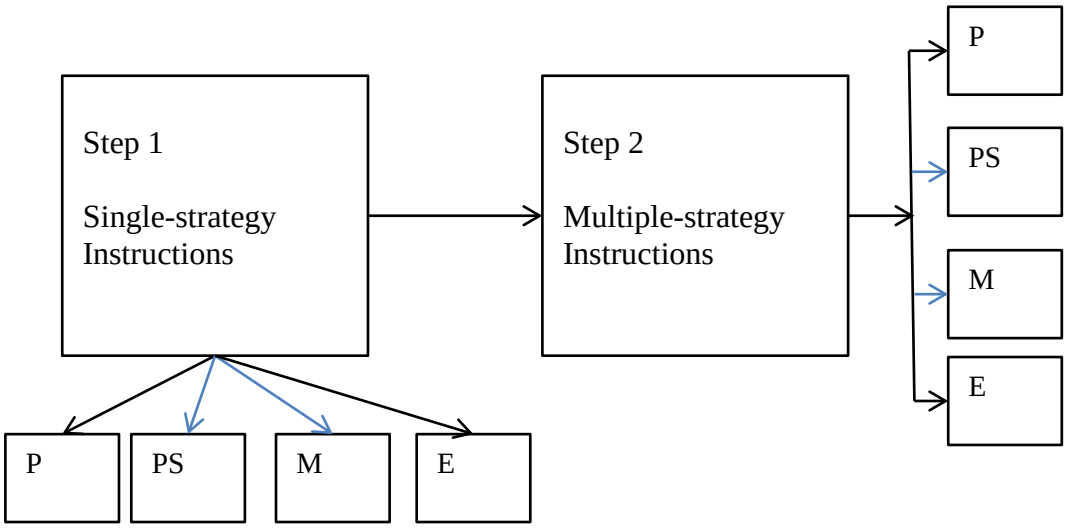

Figure 8.2 *Metacognitive instructions supported by this study*

Note. P = planning; PS = problem-solving; M = monitoring; E = evaluating.

8.5.1.2 Syllabus designing and task development

The variability in task complexity across tasks and its strong correlations with performance, individual attributes and metacognitive strategy use indicate that in designing syllabus and tasks for metacognitive instructions for speaking, EFL teachers may need to consider sequencing the speaking tasks in line with task complexity so to meet the requirements of learners who have various levels of language proficiency. Otherwise, as discussed in Chapter Seven, if tasks are too easy or too difficult inappropriately, they may negatively influence the learners' motivation and arouse anxiety. Because of this, the learners are very likely not to use their strategic competence or metacognitive strategies as they are assumed to be in performing the tasks given.

Consequently, the tasks that are designed to elicit the learners' metacognitive strategy use may not function as expected. In the end, metacognitive instructions may fail due to the tasks per se.

Furthermore, the four task complexity variables (planning time, prior knowledge, steps involved, and task type) were found to affect speaking performance at a varying degree, among which prior knowledge was the most influential. In accordance with this, EFL teachers can manipulate these factors to develop speaking tasks with different complexity for their classroom instructions (Robinson, 2005, 2011b, 2015; Robinson & Gilabert, 2020). With a better understanding of syllabus designing and tasks development, better pedagogical outcomes may be achieved in EFL classrooms for speaking, and those problems with Chinese EFL classroom instructions (see Section 1.4) may be solved. Besides, the characteristics of the integrated speaking test tasks and the relationship between the test and EFL learners' performance may inspire EFL teachers to adopt a holistic view in syllabus designing and task development by integrating reading, listening writing and speaking activities, which will assist their students to develop familiarities with the real-world language use tasks. As a result, the EFL learners' speaking ability is expected to be raised, and the wash-back effects of the integrated speaking test on speaking instructions can be accomplished (Bachman & Palmer, 2010; Newton & Nation, 2020).

8.5.2 Implications for L2 assessment

Although the test task characteristics model proposed by Bachman and Palmer (2010) is comprehensive, its applicability is hard to be achieved in the actual research efforts, which complicates the empirical studies in L2 assessment (see Section 3.3.2). Henceforth, the measurement of task complexity built upon the correspondence between Bachman and Palmer's (2010) framework and Robinson's (2015) Triadic Componential Framework (see Section 3.3.4) may provide an additional validated approach to measuring task complexity in L2 assessment. Given the under-researched situations of L2 speaking assessment and the extensive recognition of Robinson's framework, the establishment of the correspondence may simplify task complexity measurement in L2 assessment, and accordingly facilitate the endeavours along the research direction of test task validation (Bachman & Palmer, 2010).

Additionally, the research findings suggest that test developers should consider the appropriate level of the cognitive demand or task complexity imposed on test-takers. Such consideration is to ensure that the expected responses from the test-takers for examining their language ability can be prompted by the test tasks designed. As noted above, a test task that is too easy or too difficult may generate a test score that cannot truly reflect a test-takers' language ability, and in the end, the validity and reliability of the test and accordingly its usefulness may be placed into question (Bachman & Palmer, 2010; Hughes, 2017; O'Sullivan, 2013; Weir, 2005). Also, the study may inspire test developers to taken into account test-takers' relative prior knowledge, and the overweighting role of prior knowledge over the other task complexity factors in manipulating task characteristics for the overall complexity of test tasks. In this way, the assumed cognitive complexity of test tasks for various testing purposes can be achieved (Bachman & Palmer, 2010; Hughes, 2017; O'Sullivan, 2013; Wang & Yu, 2018; Weir, 2005).

8.6 Limitations

Regardless of its contributions and implications, it has to be acknowledged that this study is not without limitations, which might restrict its generalisability. The possible limitations are due to the followings.

8.6.1 Limitation of self-reported data

One limitation has to do with whether test-takers' strategic competence or their use of the metacognitive strategies can be validly elicited through self-report instruments such as questionnaires and interviews. Although self-report is widely employed in the investigation of one's internal metacognitive strategies (e.g., Qin, 2018; Zhang & Qin 2018), some researchers argue that data collected on self-reported questionnaires cannot represent the actual use of individuals' metacognitive strategies. They believe that respondents might misunderstand the items, or they might attempt to give socially desirable responses rather than providing accurate and honest responses (e.g., Cohen, 2011; Greene & Azevedo 2010; McNamara 2011; Purpupa, 1999; Schellings 2011; Veenman 2011). To increase the validity of self-report data, it is postulated that multiple procedures of data collection should be conducted (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Dörnyei, 2019). However, although the Semi-structured Interview Guide was adopted for complementing the Metacognitive Strategy Inventory, and hence

has minimised some risks to the validity (see Section 4.8.1), both were essentially self-report instruments. Due to resource constraint, diverse means were not applied in the study, which may pose a threat to the validity of the research findings.

8.6.2 Scope of sampling

As convenience sampling was employed, the population was made up of sophomores enrolled in two northern Chinese universities who had similar backgrounds, particularly, in their English learning experience and the level of their language proficiency (see Section 4.2.2). The limitations caused by the sample size and the geographic sites of the participants, therefore, may restrict the generalisation of the research results to similar populations (Daniel, 2011; Gurven, 2018; Muijs, 2011; Pallant, 2016; Teng, 2016; Zhang et al., 2014; Zhang, 2017).

8.7 Suggestions for Further Research

As discussed previously, although individual attributes were beyond the scope of this study, their mediator role in L2 assessment has been identified, which suggest the subsequent steps that might be taken in further research. In addition, the limitations presented above indicate some possible efforts in the future research of relevance. Based on these, suggestions for further studies are made in the following aspects.

First, test-takers' individual attributes should be focused on in examining the relationships among test tasks, test-takers and test performance. As reported previously, individual attributes were found working as a mediator between task complexity and strategy use, and they explained a rather large proportion of the variance in test-takers' performance (see Section 5.3.2.4). Nonetheless, constrained by the research focus, in-depth investigation of this variable was not carried out. Consequently, it was unknown what the individual attributes really were qualitatively, and how they were impacted by task complexity quantitatively, although some researchers have pointed out that task complexity affected individual attributes such as motivation and anxiety (see Section 2.5.1). Moreover, as the impact of personal attributes on test performance has not been paid appropriate attention in L2 assessment (O'Sullivan 2000, 2012), the role of individual attributes in affecting L2 speaking assessment performance warrants further quantitative and qualitative investigations.

Second, in terms of the comprehensive investigation of task complexity, more efforts in the synergetic effects of task complexity on task performance are needed with a special focus on the task variables along the dimension of resource-dispersing, and their individual and isolated impacts on task performance. Furthermore, additional work is merited in exploring task complexity and its predicative effects on performance in multiple contexts, particularly, the contexts closely related to computer-mediated language use situations when the Information Age is dominating the current real world where human beings' languages are used and researched (Chapelle & Voss, 2016; Winkle & Isbell, 2017). Likewise, as delineated in Chapter Six, some interviewees' perceptions of task complexity were influenced by fatigue generated by task sequence, distractions from testing setting and their language proficiency. These variables, though not investigated in this research, are suggested to be studied in the future in that they are associated with not only test task characteristics (e.g., the test setting, task sequence), but test-takers' factors (e.g., peer pressure, fatigue), the two fundamental factors in L2 assessment.

Third, regarding research methodology, to counterbalance instrument-related limitations, in addition to subjective self-reports, other objective means such as eye-tracking can also be employed to measure metacognitive strategies as suggested by Cohen (2014) and Oxford (2017), if conditions permit. In the similar vein, a larger sample size with more diverse backgrounds is strongly recommended for better generalisability. Sample size and sample heterogeneity have long been acknowledged as essential elements that help to produce stronger research. It is believed that the larger and the more diverse the sample is, the more accurate and more generalisable the final research outcome might be (Creswell & Creswell 2018; Creswell & Guetterman, 2019; Daniel, 2011; Gurven, 2018). It is, therefore, suggested that a larger sample size characterised by more heterogeneity is adopted in future empirical studies so that the representativeness of Chinese EFL learners or of an even larger population of EFL learners will be enhanced.

8.8 Final Thoughts as Closing Remarks

It is common that on a researcher's long and tough journey of literature review, limitations may be the most frequent word appearing on the final pages of a specific

paper or a thesis. It is also true that because of resource constraint, which often takes place in the actual research context, researchers find it impossible to harvest the expected results as the case with this study. Such impossibility may threaten the generalisability of their studies, which, to some degree, lead to the ceaselessness of the research efforts for the ultimate truth, and hence the absolute growth of the research per se. Indeed, for researchers, the suggestions for future steps often serve as the upcoming research gap to be bridged, and the final words in the final section of the conclusion chapter typically indicates a new starting point of an even longer and tougher research journey ahead. The final thoughts here, therefore, are not only the concluding reflection of my experience in conducting the study, but the preparation for the next research effort as well: “I cannot say in ‘conclusion’, because I am probably at yet another beginning, laden with some thoughts that had not occurred…” (Meloy, 2002, p. 178), and “just when I feel I’ve reached the top, I find I’m beginning all anew” (Woods, 2009, p. 18).

Appendices

Appendix 1 Metacognitive Strategies Inventory

Task Complexity, Metacognitive strategy Use, and Oral Production Performance on Integrated Speaking Tasks: A Study of Chinese EFL (English as Foreign Language) Learners

Part One

In this part please provide your information by ticking (√) in the box or write your responses in the space so we can better understand your answers.

1. Code:
2. Age:
3. Gender: Male_□ Female_□ Gender diverse_□
4. The years you have been learning English to present:

7~9 years_ □ 10~12 years_□ 13~15 years_□ Others_____.

5. English proficiency reflected by test
   CET4_□ CET6_□_ BEC_□ IELTS_□ TOEFL_□

Part Two

Please give a rating on the task you just finished. The rating should range from 1(the tasks requires no mental efforts at all; the task is not difficult at all) to 9 (the task requires extreme mental efforts; the task is extremely difficulty).

Task 3

This task requires no mental efforts at all. 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.

This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Task 4

This task requires no mental efforts at all. 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.

This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Task 5

This task requires no mental efforts at all . 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.

This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Task 6

This task requires no mental efforts at all. 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.

This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Part Three

In this part, please read each of the following statement and indicate how you thought during the integrated speaking test by ticking (√) 0 (never or almost never), 1 (rarely), 2 (sometimes), 3 (often), 4 (usually), and 5 (always or almost always)

Your thinking	0	1	2	3	4	5
1. I knew what the task questions required me to do.						
2. I was aware of the need to plan a course of action.						
3. I thought about what to do to complete the task well.						
4.I made sure I clarified the goals of the task						
5. I understood the essential steps needed to complete the task.						
6. I organized the structure of what I was going to say before speaking.						
7. I guessed the meaning of the unknown words or expressions by using my knowledge (e.g. words in the context, knowledge of word information, knowledge of the topic.						
8. I used the context to guess the topic.						
9. I drew on my background knowledge to complete the task.						
10. I made up new words or guess if I didn't know the right ones to use.						
11. I used a word or phrase that means the same thing when I could not think of a word in English.						
12.I knew when I should complete a task more quickly.						
13. I knew when I should complete a task more carefully						
14. I knew how much time had gone by.						
15. When I was speaking, I knew when I had spoken in a way that sounded like a native speaker.						
16. I related the incoming information to what I had known.						
17.When I was performing my task, I took notes on the important words and concepts.						
18. I knew what to do if my intended plan did not work efficiently during the task.						
19. I mentally give myself a grade after I finished my task.						
20. I checked whether I had accomplished my goal after completing my task.						
21. I checked the mistakes I had made in the task.						
22. I evaluated my performance satisfaction as I moved along the task.						
23. I evaluated whether my intended plans worked effectively.						

Well done! 

You have completed the questionnaire! Your help is highly appreciated!

We invite you to join us in the interviews, and if you want to join us please tell us by ticking (√) the following statement.

I want to participant the interviews_____.

中国英语学习者在任务复杂度增加的基于托福口语综合口语任务下元认知使用策略

(中文版本)

亲爱的同学们：
此次调查的目的是收集您在刚刚参加完的托福综合口语任务下所使用的元认知策略相关信息。我们希望您可以帮助我们解答关于中国学生元认知策略使用情况的相关问题。该调查问卷并非考试所以您的答案无对错之分，我们希望听到您个人的真实想法。因此我们希望您如实回答问卷以确保问卷调查的有效性。该问卷内容绝对保密。关于问卷填写人的个人信息在任何情况下均不会公开。非常感谢您的配合和合作。

第一部分

请在下面的方框中打钩(√)或在空白处填写您的回答以方便我们更好地了解您的问卷回答

1. 个人代码
2. 年龄
3. 性别： 男_ □ 女_ □
4. 到目前为止您已经学习英语学习多长时间：
   7~9 年_ □ 10~12_ □ 年 13~15 年_□ 其他---------
5. 参加过何种英语考试
   CET4_□ CET6_□ BEC_□ IELTS_□ TOEFL_□

第二部分

请对您刚刚完成的四道口语题目进行打分。1-9 表示费脑力系数（这道题一点不费脑力-这道题太费脑力）；1-9 表示难度系数（这道题一点不难-这道题太难）。

Task 3

这道题一点不费脑力。 1 2 3 4 5 6 7 8 9

这道题太费脑力。 1 2 3 4 5 6 7 8 9

这道题一点不难。 1 2 3 4 5 6 7 8 9

这道题太难了。 1 2 3 4 5 6 7 8 9

Task 4

这道题一点不费脑力。 1 2 3 4 5 6 7 8 9

这道题太费脑力。 1 2 3 4 5 6 7 8 9

这道题一点不难。 1 2 3 4 5 6 7 8 9

这道题太难了。 1 2 3 4 5 6 7 8 9

Task 5

这道题一点不费脑力。 1 2 3 4 5 6 7 8 9

这道题太费脑力。 1 2 3 4 5 6 7 8 9

这道题一点不难 。 1 2 3 4 5 6 7 8 9

这道题太难了。 1 2 3 4 5 6 7 8 9

Task 6

这道题一点不费脑力。 1 2 3 4 5 6 7 8 9

这道题太费脑力 。 1 2 3 4 5 6 7 8 9

这道题一点不难 。 1 2 3 4 5 6 7 8 9

这道题太难了。 1 2 3 4 5 6 7 8 9

第三部分

请认真阅读表格中的内容并打钩(√)选出你在综合口语考试中有哪些想法： 0 (从不或几乎不), 1 (很少), 2 (偶尔), 3 (常常), 4 (大多情况下), 5 (总是或几乎总是这样)

您的想法	**0**	**1**	**2**	**3**	**4**	**5**
1.我清楚题目要求我做什么。						
2.我明白需要规划答题过程。						
3.我想过需要做什么才能完成任务。						
4.我确信已清楚任务目标						
5.我明白完成任务所需要的主要步骤。						
6.我事先组织好了想说的内容的结构。						
7.我会借助已有的知识（如单词的上下文，构词及话题） 来猜测陌生单词或词组的意思。						
8.我根据上下文猜测话题。						
9.我借助已有知识来完成话题任务。						
10.(口语表达)想不起来某个英语单词时，我用相通常思的其他词或词组。						

11.我使用近义词之类其他方法来表达意思。						
您的想法	**0**	**1**	**2**	**3**	**4**	**5**
12.我知道答题时何时该加快速度。						
13.我知道答题中何时要更加仔细。						
14.答题过程中我知道自己用掉多少时间。						
15.口试过程中，我知道自己哪些地方说得比较地道						
16.我能把题中的信息与已有知识联系起来。						
17.答题过程中我会记录重要词语与概念。						
18.如果进展不顺利，我知道该怎么对付。						
19.完成口语任务后我在脑子里给自己的表现打了个分数。						
20.任务结束后我检查自己是否达到目标。						
21．考试后我会检查自己的错误。						
22.我会评价自己在答题表现方面的满意度。						
23.我会评价原定计划实施的有效性。						

太棒了！ 

您已经完成问卷，我们非常感谢您的参与！并真诚邀请您参加访谈，如果愿意请打钩(√)注明： 我愿意参加访谈_____。

Appendix 2 Semi-Structured Interview Guide（English Version）

1. Could you tell me what you did in the planning time for each task?

2. Could you tell me whether you had a particular goal before you started your task today?

3. Could you tell me whether you completely understood the reading and listening materials in the tests?

4. Could you tell me what you did to solve the problems in the tests if you had?

5. Could you tell me whether you had evaluated your performance in the test after the task was done?

6. You have done 2 tasks today, and which one do you think is the easiest one and which one do you think is the most difficult one? Why?

7. You have done 4 tasks in total, and which one do you think is the easiest one and which one do you think is the most difficult one? Why

8. Do you think it would be easier for you to complete the task when you were given 30s to prepare compared with 20s given? Why?

9. Do you think it would be easier for you to listen before speaking than to read and listen before speaking? Why?

10. Do you think it would be easier for you to speak on a topic related to campus life than to speak on a topic regarding an academic lecture? Why?

11. Do you think it would be easier for you to explain a concept than to providing a solution? Why?

Semi-Structured Interview Guide (Chinese Version)

i. 你能告诉我你在答题前的准备时间里都做些什么吗？

ii. 你能告诉我在开始今天的答题前你心里给自己定了什么目标吗？

iii. 你能告诉我你完全明白了今天考试中出现的阅读和听力内容吗

iv. 你能告诉我遇到问题时是怎么解决的吗（如果有的话）？

v. 你能告诉我考完试以后你给自己的表现打分了吗？

vi. 你已经完成了全部 4 个口语任务， 你认为哪一个任务最难，哪一个最简单？为什么你这么认为？

vii. 你认为 30 秒的准备时间比 20 秒准备时间更易于你完成考试任务吗？

viii. 你认为先听后说比先读、听再说更易于你完成考试任务吗？

ix. 你认为校园生活的口语任务话题比学术讲座的口语任务话题更易于你完成考试任务吗？

x. 你认为解释学术概念的口语任务话题比提供解决方案的口语任务话题更易于你完成考试任务吗？

Appendix 3 Questionnaire on Teachers' Evaluation of Task Complexity (English Version)

Dear colleagues,

The purpose of this survey is to collect information on your evaluation of task complexity. We would like to ask you to help us answer the following questions regarding the cognitive load which four integrated speaking tasks may impose on Chinese EFL students. This is not a test so there are no" right" or "wrong" answers and we are interested in your personal opinion. Please give your answers sincerely as only this guarantee the success of the investigation. The content of the form is absolutely confidential. Information identifying the respondent will not be disclosed under any circumstances.

Thank you very much for your participation and cooperation!

Part One

In this part please provide your information by ticking (√)in the box or write your responses in the space so we can better understand your answers.

6. Code:
7. Age:
8. Gender: Male□ Female□
9. The years you have been teaching English to present:

7~9 years □ 10~12 years□ 13~15 year □ Others_____.

10. English proficiency reflected by education background

Bachelor□ Master □ Doctor□

Part Two

In this part, please read each of the following statement and indicate how you thought on the task complexity of the 4 integrated speaking tasks you just performed and explain why.

Task 3

This task requires no mental efforts at all. 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.

This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Your explanation_____.

Task 4

This task requires no mental efforts at all. 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.
This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Your explanation_____.

Task 5

This task requires no mental efforts at all. 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.
This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Your explanation_____.

Task 6

This task requires no mental efforts at all. 1 2 3 4 5 6 7 8 9 This task requires extreme mental efforts.
This task was not difficult at all. 1 2 3 4 5 6 7 8 9 This task was extreme difficult.

Your explanation_____.

You have completed the questionnaire! Your help is highly appreciated!

Appendix 4 Consent Form for Student Participants

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

SCHOOL OF CURRICULUM AND PEDAGOGY

Epsom Campus
Gate 3, 74 Epsom Ave
Auckland, New Zealand
T +64 9 623 8899
W www.education.auckland.ac.nz
The University of Auckland
Private Bag 92601

CONSENT FORM
THIS FORM WILL BE HELD FOR A PERIOD OF 6 YEARS

Project: Task Complexity, Metacognitive strategy Use, and Oral Production Performance on Integrated Speaking Tasks: A Study of Chinese EFL (English as Foreign Language) Learners
Researcher: PhD candidate Weiwei Zhang
Supervisors: Professor Lawrence Zhang & Dr. Aaron Wilson

I have read the Participant Information Sheet and understood the nature of the research and why I have been invited to participate. I have had the opportunity to ask questions and have them answered satisfactorily.

- I agree to participate in the Phase Two study of this research and my participation is entirely voluntary.
- I understand that this study will be conducted out of my usual class time.
- I understand that my participation, non-participation or withdrawal will not affect my relationship with the school or my academic performance and future employment in any way.
- I understand that my responses will be audio recorded in the semi-structured interviews and I have the right to refuse to answer questions without giving any reasons at any stage of the recording.
- I understand the transcripts of the interview will be sent to me for checking and make sure that the transcriptions trace exactly what I say during the interview.
- I understand that the hard copy and digital data will be stored separately and securely for a period of six years and then destroyed.
- I understand that the data collected from the research will be used for the researcher's PhD thesis, and may be used for the academic publications and conference presentations.
- I understand that the researcher will protect my confidentiality throughout the research and no identifying information and data will be disclosed to the third party in any circumstances.
- I wish to receive a summary of findings, which can be emailed to me at this email address: _____ (If not, leave this blank.)

Name _____Signature _____ Date _____

Approved by the University of Auckland Human Participants Ethics Committee on 020972 for three years.

Appendix 5 Participant Information Sheet for Student Participants

SCHOOL OF CURRICULUM AND PEDAGOGY

Epsom Campus
Gate 3, 74 Epsom Ave
Auckland, New Zealand
T +64 9 623 8899
W www.education.auckland.ac.nz
The University of Auckland
Private Bag 92601

PARTICIPANT INFORMATION SHEET
(Students)

Project: Task Complexity, Metacognitive strategy Use, and Oral Production Performance on Integrated Speaking Tasks: A Study of Chinese EFL (English as Foreign Language) Learners
Researcher: PhD candidate Weiwei Zhang
Supervisors: Professor Lawrence Zhang & Dr. Aaron Wilson

Researcher Introduction
My name is Weiwei Zhang, a PhD student in the School of Curriculum and Pedagogy, Faculty of Education and Social Work, The University of Auckland, New Zealand. I am conducting school-based research as part of my PhD thesis.

Project description and invitation
The objective of my research is to explore the effects of degrees of task complexity on Chinese EFL learners' reported use of metacognitive strategies, oral production performance on the TOEFL-based (Test of English as Foreign Language) integrated speaking tasks, and their interrelationships.

This empirical research is a two-phase design which includes two pilot studies in Phase One and one main study in Phase Two. The Pilot studies in Phase One aim to validate the instruments, test the equipment and pilot the data scoring procedure for the main study of Phase Two. The research will be conducted out of the usual class time. Findings of this study may contribute, theoretically and pedagogically, to our current knowledge on EFL learners' metacognitive strategy use in performing increasingly complex integrated speaking tasks.
As such, you are cordially invited to join my PhD research. I have contacted your school and gained the permission to ask your involvement.

Student involvement
In Phase Two or in the main study, 400 student participants (300 participants engaging in Phase One) will be randomly assigned to four groups and they will perform 4 integrated speaking tasks (different tasks from those employed in Phase One) with 2 tasks each time. 20 minutes interval between the 2 tasks will be established to eschew the participants' fatigue and the carry-over effect. Each task will last about 5 minutes followed by 30-minute metacognitive strategies questionnaire (MSQ). After the completion of the 2 tasks, 4 participants will be selected on random basis for a 20- minute semi-structured interviews conducted on the same day. The remaining 2 tasks will be performed two days later when the same procedure in the first 2 tasks is repeated.

Participants' rights
Each student participant is entitled to withdraw without giving a reason at any time in the research. For the student participants in the semi-structure interviews they have the right to refuse to answer any specific question or leave or have the recorder turned off without giving any reasons at any stage in the interviews. Your Dean has given an assurance that your participation or non-participation will have no effect on your grades or relationship with the University. Each participant will receive a summary of the research findings upon request.

Data management
Hard copy data will be securely stored in a locked cabinet at the University of Auckland (UoA) and electronic data will be stored on a UoA password protected computer, backed up by the university server at the UoA. While in China, all the hard copy data will be locked in a cabinet of the researcher's office with security guarantee and all the electronic data will be stored in the researcher's computer which is password protected and can be accessed only by the researcher. Raters, research assistants and anyone who potentially engages in data collection and data analysis can have access to the data only with the permission of the researcher and for the purpose of the research itself. All hard copy data will be shredded and the digital information will be deleted after six years. The data collected from the research will be used for the researcher's PhD thesis at the University of Auckland, and may be used for academic publications, and conference presentations.

Benefits and risks
A copy of the summery of the research will be sent to you and each participant upon request. Participants will not incur any cost. As a token of appreciation, each participant will be given a pen at the value of 1 NZ dollar.

Anonymity and confidentiality
In this project, the researcher will ensure the identity of participants is kept confidential. When answering the questionnaires, the participants will be asked to write a unique name only recognized by them, and during data analysis, they will be given a code and a list will be used to link the participants via questionnaires and the recording files of their integrated oral performance. In case that the information provided by the participants will be reported or published for the purpose of the research, pseudonyms will be employed to protect their identities and privacy. No identifying information or data obtained from the study will be disclosed to a third party for any purposes under any conditions.

Contact Details and Approval
Researcher contact details
Weiwei Zhang PhD student in the School of Curriculum and Pedagogy
Faculty of Education and Social Work
University of Auckland
Email: wzha589@auckland.ac.nz
Local contacts in China: 0086 13155250017 (cell phone);
swwanbei1@ swliuxue.org (E-mail).

Main Supervisor

Professor Lawrence Zhang
School of Curriculum and Pedagogy
Faculty of Education and Social Work
University of Auckland
Phone: +64 9 373 7999, ext 48750,
Email: lj.zhang@auckland.ac.nz

Co-supervisor

Dr. Aaron Wilson
School of Curriculum and Pedagogy
Faculty of Education and Social Work, University of Auckland
Phone: + 64 93737999 ext 48574,
Email: aj.wilson@auckland.ac.nz

Head of School

Helen Hedges
School of Curriculum and Pedagogy
Faculty of Education and Social Work
University of Auckland
Phone: +64 9 373 7599, ext 48606,
Email: h.hedges@auckland.ac.nz

For any queries regarding ethical concerns you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Research Office, Private Bag 92019, Auckland 1142. Telephone 09 373-7599 ext. 83711.
Email: ro-ethics@auckland.ac.nz.

Approved by the University of Auckland Human Participants Ethics Committee on 020972 for three years.

References

Abdi Tabari, M. (2020). Differential effects of strategic planning and task structure on L2 writing outcomes. *Reading & Writing Quarterly, 36*(4), 320-338.

Abdolrezapour, P. (2019). Applying computer-mediated active learning intervention to improve L2 listening comprehension. *Applied Research on English Language, 8*(4), 511-530.

Adams, R., & Alwi, N. A. N. M. (2014). Prior knowledge and second language task production in text chat. In M. González-Lloret, & L. Ortega (Eds.). *Technology-mediated TBLT: Researching technology and tasks* (Vol. VI, pp. 51-78). Philadelphia, PA: Benjamins.

Ajzen, I. (2002). Perceived behavioural control, self-efficacy, locus of control, and the theory of planned behaviour. *Journal of Applied Social Psychology, 32*(4), 665-683.

Akkakoson, S. (2013). The relationship between strategic reading instruction, student learning of L2-based reading strategies and L2 reading achievement. *Journal of Research in Reading, 36*(4), 422-450.

Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing, 26*(4), 621-6.

Allen, M. (2017). *The SAGE encyclopaedia of communication research methods*. Los Angeles, CA: Sage.

Almasi, J. F., & Fullerton, S. K. (2012). *Teaching strategic processes in reading* (2nd ed.). New York, NY: Guilford.

Amani, S. (2014*). Metacognitive Strategy Instruction and Pre-Task Planning: Impact on L2 Argumentative Writing Ability*. Unpublished Ph.D. thesis, The University of Auckland, Auckland, New Zealand.

An, Y., & Cao, L. (2014). Examining the effects of metacognitive scaffolding on students' design problem solving and metacognitive skills in an online environment. *Journal of Online Learning and Teaching, 10*(4), 552-568.

Anderson, N. J. (2002). The role of metacognition in second language teaching and learning. ERIC ED463659. Retrieved from https://eric.ed.gov/?id=ED463659

Anderson, N. J. (2008). Metacognition and the good language learner. In C. Griffiths (Ed.), *Lessons from good language learners* (pp. 99–109). Cambridge, England: Cambridge University Press.

Anderson, N. J. (2012). Metacognition: Awareness in language learning. In S. Mercer, S. Ryan, & M. Williams (Eds.), *Psychology for language learning: Insights from research, theory and pedagogy* (pp. 169–187). London, England: Palgrave.

Aydin, S. (2013). Factors affecting the level of test anxiety among EFL learners at elementary schools. *E-international Journal of Educational Research, 4*(1), 63-81.

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning, 4*(1), 87-95.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge, England: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal, 2*(1), 1-34.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41-71). Ottawa: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer A. S. (2010*). Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford, England: Oxford University Press.

Bagarić, V., & Mihaljević Djigunović, J. (2007). Defining communicative competence. *Metodika, 8,* 94–103.

Bai, B., & Guo, W. (2019). Motivation and self-regulated strategy use: Relationships to primary school students' English writing in Hong Kong. *Language Teaching Research,* 1-22. Online first. doi:10.1177/1362168819859921

Baralt, M. (2010). *Task complexity, the Cognition Hypothesis, and interaction in CMC and FTF environments.* Unpublished Ph.D. thesis. Georgetown University, Georgetown, Washington, DC, USA.

Baralt, M. (2014). Task complexity and task sequencing in traditional versus online language classes. In M. Baralt, R. Gilabert, & P. Robinson (Eds.), *Task sequencing and instructed second language learning* (pp.95-122). London, England: Bloomsbury.

Barkaoui, K. (2010a). Explaining ESL essay holistic scores: A multilevel modelling approach. *Language Testing, 27*(4), 515-535.

Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly: An International Journal, 7*(1), 54-74.

Barkaoui, K. (2013). Using multilevel modelling in language assessment research: A conceptual introduction. *Language Assessment Quarterly: An International Journal, 10*(3), 241-273.

Barkaoui, K. (2015). *Test-takers' writing activities during the TOEFL iBT® writing tasks: A stimulated recall study* (Research Report No. RR-15-01), Princeton, NJ: Educational Testing Service.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviours in independent and integrated speaking tasks. *Applied Linguistics, 34*(3), 304-324.

Bazeley, P. & Jackson K. (2013). *Qualitative data analysis with* NVivo (2nd ed.). London, England: Sage.

Beatty, P. C., Collins, D., Kaye, L., Padilla, J., Willis, Gordon B., & Wilmot, A. (2020). *Advances in questionnaire design, development, evaluation and testing*. Newark, NJ: Wiley.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation, 18*(6), 1-13.

Bell, D. J., & Ruthven, L. (2004). Searcher's assessments of task complexity for web searching. In S. McDonald, & J. Tait (Eds.), *Advances in information retrieval: ECIR lecture notes in computer Science* (pp. 57-71). Heidelberg, Berlin: Springer.

Berger, J., & Karabenick, S. A. (2011). Motivation and students' use of learning strategies: Evidence of unidirectional effects in mathematics classrooms. *Learning and Instruction, 21*(3), 416-428.

Borkowski, J. G., Chan, L. K. S., & Muthukrishna, N. (2000). A process-oriented model of metacognition: Links between motivation and executive functioning. In G. Schraw, & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 1-41). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT™ speaking test scores as indicators of oral communicative language proficiency. *Language Testing 29*(1), 91-108.

Brinkmann, S., & Kvale, S. (2015). *Interviews: Learning the craft of qualitative research interviewing*. Los Angeles, CA: Sage.

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT™ and real-life academic speaking activities. *Language Assessment Quarterly: An International Journal, 11*(4), 353-373.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert, & R. H. Kluwe (Eds.) *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Erlbaum.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J.H. Flavell & EM. Markman (Eds.), *Cognitive development* (Vol. III of P.H. Mussen, Ed., *Handbook of child psychology*, pp. 77-166). New York, NY: Wiley.

Brown, A., & Ducasse, A. M. (2019). An equal challenge? comparing TOEFL iBT™ speaking tasks with academic speaking tasks. *Language Assessment Quarterly: An International Journal, 16*(2), 253-270.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (Research Report No. RR-05-01). Princeton, NJ: Educational Testing Service.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford.

Bygate, M. (1987). *Speaking.* Oxford, England: Oxford University Press.

Bygate, M. (2011). Teaching and testing speaking in M. H. Long, & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 412-440). West Sussex, England: Wiley.

Byrne, B. M. (2016). *Structural equation modelling with AMOS Basic concepts, applications, and program* (3rd ed.), New York, NJ: Routledge.

Cai J. (2012). *Future direction for Chinese EFL education at tertiary Level*. Shanghai: Shanghai Jiaotong University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. *Language and Communication, 1*(1), 1-47.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics,1*(1), 1-47.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In: *Testing the English proficiency of foreign students*. Washington, DC: Centre for Applied Linguistics. Reprinted in H. B. Allen, & R. N. Campbell (Eds.). (1972), *Teaching English as a second language: A book of readings* (pp. 313– 320). New York, NY: McGraw Hill.

Chamot, A., Barnhardt, S., Beard El- Dinary, P., Robbins, J. (1999). *The learning strategy handbook.* New York, NY: Longman.

Chamot, A. U. (2005). Language learning strategy instruction: Current issues and research. *Annual Review of Applied Linguistics*, *25*, 112-130.

Chamot, A. U. (2009). *The CALLA handbook: Implementing the cognitive academic language learning approach (*2nd ed.). White Plains, NY: Pearson Education.

Chamot, A. U., & Harris, V. (Eds.). (2019). *Learning strategy instruction in the language classroom: Issues and implementation*. Bristol, England: Multilingual Matters.

Chang, L. (2012). *Investigating the relationships between Chinese university EFL learners' metacognitive listening strategies and their comprehension and incidental vocabulary acquisition from listening tasks.* Unpublished Ph.D. thesis, The University of Auckland, Auckland, New Zealand.

Chang, C. H., & Liu, H. J. (2013). Language learning strategy use and language learning motivation of Taiwanese EFL university students. *Electronic Journal of Foreign Language Teaching*, *10*(2), 196-209.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In F. Bachman, & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing* (pp. 32-70). New York, NY: Cambridge University Press.

Chapelle, C. A. (2009). The relationship between second language acquisition theory and Computer-Assisted language learning. *Modern Language Journal, 93*(1), 741-753.

Chapelle, C. A. (2010). The spread of computer-assisted language learning. *Language Teaching*, *43*(1), 66-74.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a validity argument for the test of English as a foreign language.* New York, NY: Routledge.

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning Technology, 20*(2), 116-128.

Chen, C. W. Y. (2019). Guided listening with listening journals and curated materials: A metacognitive approach. *Innovation in Language Learning and Teaching, 13*(2), 133-146.

Chen H. C. (1990). Lexical processing in a non-native language: Effects of language proficiency and learning strategy. *Memory & Cognition,18*(3), 279-288.

Chen, J., & Ed, M. (2007). *Case study of the TOEFL iBT preparation course.* Unpublished Ph.D. thesis. The University of Western Ontario, West Ontario, Canada.

Chou, M. (2013). Strategy use for reading English for general and specific academic purposes in testing and non-testing contexts. *Reading Research Quarterly, 48*(2), 175-197.

Cohen, A. D. (2014). *Strategies in learning and using a second language*. London, England: Routledge.

Cohen, A. D. (2018). Moving from theory to practice: A close look at language learner strategies in R. L., Oxford, &, C. M., Amerstorfer (Eds.), *Language learning strategies and individual learner characteristics: Situating strategy use in diverse contexts* (pp. 31-54). London, England: Bloomsbury.

Cohen, A. D., & Macaro, E. (Eds.). (2007). *Language learner strategies: Thirty years of research and practice.* Oxford, England: Oxford University Press.

Cohen, A. D., Oxford, R. L., & Chi, J. C. (2002). *Learning style survey: Assessing your own learning styles.* Minneapolis, MN: Centre for Advanced Research on Language Acquisition. University of Minnesota. Retrieved from http://www.carla.umn.edu.

Corriero, E, F. (2017). Counterbalancing in M., Allen, (ed), *The sage encyclopaedia of communication research methods* (pp. 278-281). Los Angeles, CA: Sage.

Cox, T. L. (2013). *Investigating prompt difficulty in an automatically scored speaking performance assessment.* Unpublished Ph.D. thesis. Brigham Yong University, Hawaii, USA.

Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning, 15*(2), 155-213.

Cramer, D., & Howitt, D. L. (2004). *The Sage dictionary of statistics: a practical resource for students in the social sciences*. London, England: Sage.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Los Angeles, CA: Sage.

Creswell, J. W., & Guetterman, T. C., (2019). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). New York, NY: Pearson.

Creswell. J. W., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches* (4th ed.). Los Angeles, CA: Sage.

Crossley, S. A., & Kim, Y. (2019). Text integration and speaking proficiency: Linguistic, individual differences, and strategy use considerations. *Language Assessment Quarterly: An International Journal, 16*(2), 217-235.

Crystal, D. (2011). *A dictionary of linguistics and phonetics* (6th ed.). London, England: Blackwell.

Dabarera, C., Renandya, W. A., & Zhang, L. J. (2014). The impact of metacognitive scaffolding and monitoring on reading comprehension. *System, 42 (1),* 462-473.

Daniel, J. N. (2011). *Sampling essentials: Practical guidelines for making sampling choices.* Los Angeles, CA: Sage.

Davies, A. (2013). Fifty years of language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–19). Hoboken, NJ: Wiley-Blackwell.

Davis, L. (2018). Analytic, holistic, and primary trait marking scales. In J. I. Liontas, A. Shehadeh (Eds.), *The TESOL encyclopaedia of English language teaching, Vol. II: Approaches and methods in English for speakers of other languages* (Vol. II, pp.1-6). Hoboken, NJ: Wiley-Blackwell.

Dawadi, S. (2017). The relationship between reading strategy use and EFL test performance. *Journal of NELTA, 22*(1-2), 38–52.

de Boer, H., Donker, A. S., Kostons, D. D., & van der Werf, G. P. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review, 24*, 98-115.

Deci, E. L., & Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behaviour. *Psychological Inquiry, 11*(4), 227-268.

Declerck, M., & Kormos, J. (2012). The effect of dual task demands and proficiency on second language speech production. Bilingualism: *Language and Cognition, 15*(4), 782-796.

DeVellis, R. F. (2012*). Scale development: Theory and applications* (3rd ed.). Los Angeles, CA: Sage.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Erlbaum.

Dörnyei, Z. (2019). From integrative motivation to directed motivational currents: The evolution of the understanding of L2 motivation over three decades. In M., Lamb, K., Csizér, A., Henry, & Ryan, S. (Eds.). *The Palgrave handbook of motivation for language learning* (pp. 39-69). Cham, Switzerland: Palgrave.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition, 20* (3), 349-385.

Dörnyei, Z., Muir, C., & Ibrahim, Z. (2014). Directed motivational currents. In D., Lasagabaster, A., Doiz., & J. M., Sierra (Eds.), *Motivation and foreign language learning: From theory to practice* (pp. 9-30). Amsterdam, the Netherlands: Benjamins.

Dörnyei, Z., & Scott, M. L. (1997). Communication strategies in a second language: Definitions and taxonomies. *Language Learning*, *47*(1), 173-210.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. New York, NY: Routledge.

Dou S.F. (2013). 浅谈中职学校学生英语会话能力的提高[Examination on how to improve the Chinese EFL learners' speaking competence in vocational education. *Information on Science and Technology*, *17*, 275-275.

Drisko, J. W., & Maschi, T. (2015). *Content analysis: Pocket guides to social work.* Oxford, England: Oxford University Press.

Efklides, A. (2002). The systemic nature of metacognitive experiences. In P. E. Chambres, M. E. Izaute, & P. J. E. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 19-34). Boston, MA: Springer.

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review, 1*(1), 3-14.

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277-287.

Ekhlas, N. N., & Shangarffam, N. (2013). The relationship between determinant factors of self-regulation strategies and main language skills and overall proficiency. *Procedia - Social and Behavioural Sciences, 70*, 137-147.

Elder, C., Iwashita, N., & Mcnamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347-368.

Elfi, E. (2016). A study of students' planning of metacognitive strategies in listening comprehension. Ta'Dib,*17*(1), 66-74.

Ellis, R. (1984a). *Classroom second language development.* Oxford, England: Oxford University Press.

Ellis, R. (1984b). Communicative strategies and the evaluation of communicative performance. *English Language Teaching Journal*, *38* (1), 39–44.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, England: Oxford University Press.

Ellis, R. (2017). Position paper: Moving task-based language teaching forward. *Language Teaching*, *50*(4), 507-526.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). *Task-based language teaching: Theory and practice*. Cambridge, England: Cambridge University Press.

Ergai, A., Cohen, T., Sharp, J., Wiegmann, D., Gramopadhye, A., & Shappell, S. (2016). Assessment of the human factors analysis and classification system (HFACS): Intra-rater and inter-rater reliability. *Safety Science, 82*, 393-398.

Educational Testing Service. (2019). *TOEFL iBT®Test and Score Data Summary 2019.* Princeton, NJ: Educational Testing Service.

Fang, F. (2018). Ideology and identity debate of English in China: Past, present and future. *Asian English, 20*(1), 15-26.

Farhady, H. (2018). History of language testing and assessment. In J. I. Liontas & M. DelliCarpini (Eds.), *The TESOL Encyclopaedia of English Language Teaching* (pp.1-7). Hoboken, NJ: Wiley-Blackwell.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly: An international Journal 10*(3), 274-291.

Farr, F., Murray, L., & Taylor & Francis. (2016). *The Routledge handbook of language learning and technology*. London, England: Routledge.

Fazilatfar, A. M. (2010). A study of reading strategies using task-based strategy assessment. *Journal of English Language Teaching and* Learning, *53*(217), 20-44.

Fernandez, C. (2018). Behind a spoken performance: Test-takers' strategic reactions in a simulated part 3 of the IELTS speaking test. *Language Testing in Asia, 8*(1), 1-20.

Field, A. P. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Los Angeles, CA: Sage.

Finkbeiner, C., Knierim, M., Smasal, M., & Ludwig, P. H. (2012). Self-regulated cooperative EFL reading tasks: Students' strategy use and teachers' support. *Language Awareness, 21*(1-2), 57-83.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906-911.

Flavell, J. H. (1981) Cognitive monitoring. In P. Dickson (Ed.), *Children's oral communication skills* (pp 35–60). New York, NY: Academic.

Flavell, J. (2000, June). *Developing intuitions about the mental experiences of self and others*. Paper presented at the Meeting of the Jean Piaget Society, Montreal, Canada.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second language acquisition*, *18*(3), 299-323.

Fox, J., Wesche, M., Bayliss, D., Cheng, L., Turner, C.E. & Doe, C. (2007). *Language testing reconsidered*. Ottawa: University of Ottawa Press.

Frey, B. B. (2018). *The sage encyclopaedia of educational research, measurement, and evaluation.* Los Angeles, CA: Sage.

Fulcher, G. (2012a). Scoring performance tests. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–392). London, England: Routledge.

Fulcher, G. (2012b). Interlocutor and rater training. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–392). London, England: Routledge.

Fulcher, G. (2014). *Testing second language speaking.* London, England: Routledge.

Fulcher, G. (2015a). Assessing second language speaking. *Language Teaching*, *48*(2), 198-216.

Fulcher, G. (2015b). *Re-examining language testing: A philosophical and social enquiry*. London, England: Routledge.

Fulcher, G. (2018). Assessing spoken production In J. I. Liontas, A. Shehadeh (Eds.) *The TESOL encyclopaedia of English language teaching* (Vol. II, pp. 682-792). Hoboken, NJ: Wiley-Blackwell.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. New York, NY: Routledge.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344.

Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Language Assessment Quarterly: An International Journal, 9*(2), 133-151.

Garson, G. D. (2013). *Hierarchical linear modelling: Guide and applications.* Los Angeles, CA: Sage.

Ghapanchi, Z., & Taheryan, A. (2012). Roles of linguistic knowledge, metacognitive knowledge and metacognitive strategy use in speaking and listening proficiency of Iranian EFL learners. *World Journal of Education, 2*(4), 64-75.

Gilabert, R. (2005). *Task complexity and L2 narrative oral production.* Unpublished Ph.D. thesis. The University of Barcelona, Barcelona, Spain.

Gilabert, R. (2007a). The simultaneous manipulation of task complexity along planning time and [+/-Here-and-Now]: Effects on L2 oral production. In M. P. Garcia Mayo (ed.), *Investigating tasks in formal language learning* (pp. 44-68). Clevedon, England: Multilingual Matters.

Gilabert, R. (2007b). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching, 45*(3), 215-240.

Gilabert, R., Barón, J., & Llanes, À. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *International Review of Applied Linguistics in Language Teaching, 47*(3-4), 367-395.

Goh, C. (2008). Metacognitive instruction for second language listening development: Theory, practice and research implications. *RELC Journal, 39*(2), 188-213.

Gorsuch, G. (2011). *Exporting English pronunciation from China: The communication needs of young Chinese scientists as teachers in higher education abroad*. Forum on Public Policy 2011, 3. retrieved from http://forumonpublicpolicy.com/vol2011no3/archive/gorsuch.pdf

Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist, 45*(4), 203-209.

Griffiths, C. (2013). *The strategy factor in successful language learning.* Bristol, England: Multilingual Matters.

Griffiths, C. (2020). Language Learning Strategies: Is the Baby Still in the Bathwater? *Applied Linguistics, 41(*4), 607–611.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31*(1), 111-133.

Gu P-Y (2019) Approaches to learning strategy instruction. In: A.U. Chamot, & V. Harris (eds), *Learning strategy instruction in the language classroom: Issues and implementation* (pp. 22-37). Bristol, England: Multilingual Matters.

Gu, P. Y. (2003). Vocabulary learning in a second language: Person, task, context and strategies. *TESL-EJ, 7*(2), 1-25.

Gurven, M. (2018). Broadening horizons: Sample diversity and socioecological theory are essential to the future of psychological science. *Proceedings of the National Academy of Sciences of the United States of America, 115*(45), 11420–11427. Retrieved from https://doi.org/10.1073/pnas.1720433115.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Gaithersburg, MD: Advanced Analytics, LLC.

Haidar, S., & Fang, F. (2019). English language in education and globalization: A comparative analysis of the role of English in Pakistan and China. *Asia Pacific Journal of Education, 39*(2), 165-176.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis: Pearson new international edition*. Essex, England: Pearson Education.

Hair, J. F. (2019). *Multivariate data analysis* (8th ed.). Andover, Hampshire: Cengage.

Halliday, M. A. K., & Matthiessen, C. M. (2013). *Halliday's introduction to functional grammar* (4 th ed.). New York, NY: Routledge.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Harputlu, L., & Ceylan, E. (2014). The effects of motivation and metacognitive strategy use on EFL listening proficiency. *Procedia Social and Behavioural Sciences, 158*, 124-131.

Harris, L. R., & Brown, G. T. (2010). Mixing interview and questionnaire methods: Practical problems in aligning data. *Practical Assessment, Research, and Evaluation, 15*(1), 1-14.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (7th ed.). New York, NY: Guilford.

Hidri, S. (2018). Introduction: State of the art of assessing second language abilities. In S., Hidri (Ed.), *Revisiting the assessment of second language abilities: From theory to practice* (pp. 1-19). Cham, Switzerland, Springer,

Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics, 21*(1), 112-126.

Horvath, M., Herleman, H. & Lee McKie, R. (2006). Goal orientation, task difficulty, and task interest: A multilevel analysis. *Motiv Emot, 30*, 169–176.

Huang, H. D. (2018). Modelling the relationships between anxieties and performance in second/foreign language speaking assessment. *Learning and Individual Differences, 63*, 44-56.

Huang, H. D., & Hung, S. A. (2013). Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly, 47*(2), 244-269.

Huang, H. D., Hung, S. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27-49.

Huang, H. T. D. (2016). Exploring strategy use in L2 speaking assessment. *System, 63*, 13-27.

Huang, L. (2013). Cognitive processes involved in performing the IELTS speaking test: Respondents' strategic behaviours in simulated testing and non-testing contexts. *IELTS Research Reports Online Series No.6* (PP.1-51). British Council and IDP Australia.

Hughes, R. (2017). *Teaching and researching speaking.* Harlow, England*:* Longman.

Hughes, R., & Reed. B. S. (2017). *Teaching and researching speaking* (3rd ed.). New York, NY: Routledge.

Huta, V. (2014). When to use hierarchical linear modelling. *Tutorials in Quantitative Methods for Psychology, 10*(1), 13-28.

Hymes, D. H. (1972). On Communicative Competence. In J. B., Pride, & J., Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Baltimore, MD: Penguin.

Ikeda, M., Takeuchi, O. (2000). Tasks and strategy use: Empirical implications for questionnaire studies. *JACET Bulletin, 31*, 21–32.

Ikeda M, Takeuchi O (2003) Can strategy instruction help EFL learners to improve their reading ability? An empirical study. *JACET Bulletin, 37*, 49–60.

Isbell, D., & Winke, P. (2019). Test review: ACTFL Oral Proficiency Interview-Computer ® (OPIc). *Language Testing, 36*(3), 467-477.

Iwashita, N., Mcnamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an Information-Processing approach to task design. *Language Learning, 51*(3), 401-436.

Jacobse, A., & Harskamp, E. (2012). Towards efficient measurement of metacognition in mathematical problem solving. *Metacognition and Learning, 7*(2), 133-149.

Jamieson, J., & Poonpon, K. (2013b). *Developing analytic rating guides for TOEFL iBT's integrated speaking tasks* (Research Report No. RR-13-01). Princeton, NJ: Educational Testing Service.

Jamieson-Noel, D., & Winne, P. H. (2003). Comparing self-reports to traces of studying behaviour as representations of students' studying and achievement. *German Journal of Educational Psychology, 17*(3/4), 159-171.

Julkunen, K. (1989). *Situation- and task-specific motivation in foreign language learning and teaching.* Unpublished Ph.D. thesis. University of Joensuu, Joensuu, Finland.

Jumaat, N. F., & Tasir, Z. (2016). A framework of metacognitive scaffolding in learning authoring system through Facebook. *Journal of Educational Computing Research, 54*(5), 619-659.

Kahan, D. (2010). Fixing the communications failure. *Nature, 463*(7279), 296-297.

Kaivanpanah, S., Yamouty, P., & Karami, H. (2012). Examining the effects of proficiency, gender, and task type on the use of communication strategies. *Porta Linguarum, 17*, 79-94.

Karatas, H., Alci, B., & Aydin, H. (2013). Correlation among high school senior students' test anxiety, academic performance and points of university entrance exam. *Educational Research and Reviews, 8*(13), 919-926.

Karazsia, B. T., & Berlin, K. S. (2018). Can a mediator moderate? Considering the role of time and change in the mediator-moderator distinction. *Behaviour Therapy, 49*(1), 12-20.

Keats, D. M. (2000). *Interviewing: A practical guide for students and professionals*. Sydney, NSW: UNSW Press.

Khan, S. (2010). *Strategies and spoken production on three oral communication tasks a study of high and low proficiency EFL learners.* Unpublished Ph.D. thesis. The Autonomous University of Barcelona, Cerdanyola del Vallè, Spain.

Kim, H. (2009). *Investigating the effects of context and task type on second language speaking ability.* Unpublished Ph.D. thesis. Columbia University, Columbia, USA.

Kim, H., & Millsap, R. (2014). Using the bollen-stine bootstrapping method for evaluating approximate fit indices. *Multivariate Behavioural Research, 49*(6), 581-596.

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory: L2 question development through recasts in a laboratory setting. *Studies in Second Language Acquisition, 37*(3), 549-581.

Kimberlin, L., C., & Winterstein, G., A. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276-2284.

Kline, R. B. (2016). *Principles and practice of structural equation modelling* (4th ed.). New York, NY: Guilford.

Kluwe, R. H. (1987). Executive decisions and regulation of problem-solving behaviour. In F. Weinert, & R. Kluwe (Eds.), *Metacognition. motivation. and understanding* (pp. 31-64). Hillsdale, NJ: Erlbaum.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Erlbaum.

Kormos, J. (2011). Speech production and the Cognition Hypothesis. In P. Robinson (Ed.*), Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*. Amsterdam, the Netherland: Benjamins.

Kormos, J., & Wilby, J. (2019). Task motivation. In M. Lamb, K. Csizér, A. Henry, & S. Ryan (Eds.) *The Palgrave handbook of motivation for language learning* (pp. 267-286). Cham, Switzerland: Palgrave.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics in Language Teaching, 45*(3), 261-284.

Kyle, K., Crossley, S. A., & Mcnamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319-340.

Lado, R. (1961*). Language testing: The construction and use of foreign languages tests: A teacher's book*. London, England: Longman.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition, 39*(1), 167-196.

Lee, J. (2018). *The interactive effects of task complexity, task condition, and cognitive individual differences on L2 writing.* Unpublished Ph.D. thesis. University of Maryland, Maryland, USA.

Lee, J. (2019). Task complexity, cognitive load, and L1 speech. *Applied linguistics, 40*(3), 506-539.

Levelt, W. J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levkina, M. (2008). *The effects of increasing cognitive task complexity along [+/- planning Time] and [+/ – few Elements] on L2 oral production.* Unpublished Master thesis. University of Barcelona, Barcelona, Spain.

Levkina, M., & Gilabert, R. (2012). The effects of cognitive task complexity on L2 oral production. In A. Housen, F. Kuiken, & I. Vedder (Eds.) *Dimensions of L2 performance and proficiency investigating complexity, accuracy, and fluency in SLA* (pp. 171-198). Philadelphia, PA: Benjamins.

Lien, H. (2016). Effects of EFL individual learner variables on foreign language reading anxiety and metacognitive reading strategy use. *Psychological Reports, 119*(1), 124-135.

Liu, M. (2006). Anxiety in Chinese EFL students at different proficiency levels. *System, 34*(3), 301-316.

Liu, M. (2018). Interactive effects of English-speaking anxiety and strategy use on oral English test performance of high-and low-proficient Chinese university EFL learners. *Cogent Education, 5*(1), 1-14.

Liu, P., & Li, Z. (2011). Toward understanding the relationship between task complexity and task performance In *International Conference on Internationalization, Design and*

Global Development (pp.192-200). Heidelberg, Berlin: Springer. Retrieved from doi:10.1007/978-3-642-21660-2_22

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics, 42*(6), 553-568.

Livingston, J. A. (2003). Metacognition: An Overview. ERIC ED474273. Retrieved from https://eric.ed.gov/?id=ED474273.

Li, W., Lee, A., & Solmon, M. (2007). The role of perceptions of task difficulty in relation to self-perceptions of ability, intrinsic value, attainment value, and performance. *European Physical Education Review,13*(3), 301-318.

Loizidou, A., & Koutselini, M. (2007). Metacognitive monitoring: An obstacle and a key to effective teaching and learning. *Teachers and Teaching: Theory and Practice, 13*(5), 499-519.

Lu, Z., & Liu, M. (2011). Foreign language anxiety and strategy use: A study with Chinese undergraduate EFL learners. *Journal of Language Teaching and Research, 2*(6), 1298-1305.

Luoma, S. (2004). *Assessing speaking.* Cambridge, England: Cambridge University Press.

Lynch, R., Hurley, A., Cumiskey, O., Nolan, B., & McGlynn, B. (2019). Exploring the relationship between homework task difficulty, student engagement and performance. *Irish Educational Studies, 38*(1), 89-103.

Ma, J. H. (2009). *Autonomy, competence, and relatedness in L2 learners' task motivation: A self-determination theory perspective.* Unpublished Ph.D. thesis. University of Hawai'i, Hawai'i, USA.

Macaro, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *Modern Language Journal, 90*(3), 320-337.

Mackinnon, D. P. (2011). Integrating mediators and moderators in research design. *Research on Social Work Practice, 21*(6), 675-681.

Madarsara, M. A., & Rahimy, R. (2015). Examining the effect of task complexity and sequence on speaking ability of Iranian EFL learners. *International Journal of Applied Linguistics and English Literature, 4*(1), 247-254.

Mahdavi, M. (2014). An overview: Metacognition in education. *International Journal of Multidisciplinary and Current Research, 2*(6), 529-535.

Martínez, J. J. R., Pérez, M. L. V., & Navarrete, J. H. (2016). Language learning strategy use by Spanish EFL students: The effect of proficiency level, gender, and motivation. *Revista de Investigación Educativa, 34*, 133–149.

Masrom, U. K., Alwi, N. A. N. M., & Daud, N. S. M. (2015). The effects of task complexity on the complexity of the second language written production. *Journal of Second Language Teaching & Research, 4*(1), 38-66.

May, S. (2017). *Encyclopaedia of language and education* (3rd ed.). Cham, Switzerland: Springer.

McCormick, C. B. (2003). Metacognition and learning. In W. Reynolds, M. Weiner, & G.E. Miller (Eds.), *Handbook of psychology* (pp. 79-102). Hoboken, NJ: Wiley-Blakwell.

McNamara, D. (2011). Measuring deep, reflective comprehension and learning strategies: Challenges and successes. *Metacognition and Learning, 6*(2), 195-203.

McNamara, T. F. (1996). *Measuring second language performance*. London, England: Longman.

McNamara, T. F. (2000). *Language testing.* Oxford, England: Oxford University Press.

Meissel, K. L. D. (2014). *Quantitative analysis selection in education: Potential impacts on researchers' conclusions.* Unpublished Ph.D. thesis. The University of Auckland, Auckland, New Zealand.

Meijer, J., Sleegers, P., Elshout-Mohr, M., Daalen-Kapteijns, M. v., Meeus, W., & Tempelaar, D. (2013). The development of a questionnaire on metacognition for students in higher education. *Educational Research, 55*(1), 31-52.

Meloy, J. M. (2002). *Writing the qualitative dissertation: Understanding by doing* (2nd ed.). Mahwah, NJ: Erlbaum.

Ministry of Education. (2017). *A Guide for University English Teaching*. Retrieved from https://flc.ahnu.edu.cn/info/1044/6562.htm

Muijs, D. (2011). *Doing quantitative research in education with SPSS* (2nd ed.). London, England: Sage.

Namaziandost, E., & Ahmadi, S. (2019). The assessment of oral proficiency through holistic and analytic techniques of scoring: A comparative study. *Applied Linguistics Research Journal, 3*(2), 70-82.

National Capital Language Resource Centre (NCLRC). (2003). *Elementary immersion learning strategies resource guide*. Washington, DC: National Capital Language Resource Centre.

National Education Examinations Authorities（2020）. *How to interpret scores for College English Test*. Beijing, China. Retrieved from http://cet.neea.edu.cn/html1/folder/19081/5124-1.htm

Nazarieh, M. (2016). A brief history of metacognition and principles of metacognitive instruction in learning. *Journal of Humanities, Arts, Medicine and Sciences (BEST: JHAMS), 2*(2), 61-64.

Neary-Sundquist, C. A. (2008). The role of task type and proficiency level in second language speech production. Unpublished Ph.D. thesis. Purdue University, West Lafayette, USA.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*(2), 102-116.

Nett, U. E., Goetz, T., Hall, N. C., & Frenzel, A. C. (2012). Metacognitive strategies and test performance: An experience sampling analysis of students' learning behaviour. *Education Research International, 2012,* 1-16.

Newton, J. M., & Nation, I. S. P. (2020). *Teaching ESL/EFL listening and speaking* (2nd. Ed.). New York, NY: Routledge.

Nezlek, J. B. (2008). An introduction to multilevel modelling for social and personality psychology. *Social and Personality Psychology Compass*, *2*(2), 842-860.

Nezlek, J. B. (2011). *Multilevel modelling for social and personality psychology* (1st ed.). London, England: Sage.

Nishishiba, M., Jones, M., & Kraner, M. (2014). Sample selection. In M. Nishishiba, M. Jones, & M, Kraner (Eds.) *Research methods and statistics for public and non-profit administrators: A practical guide* (pp.74-85). London, England: Sage.

Nourdad, A., & Ajideh, P. (2019). On the relationship between test-taking strategies and EFL reading performance. *Journal of English Language Teaching and Learning, 11*(23), 189-219.

O'Grady, S. (2018). Investigating the use of an empirically derived, binary-choice and boundary-definition (EBB) scale for the assessment of English language spoken proficiency. In S. Hidri (Ed.). *Revisiting the assessment of second language abilities: From theory to practice* (pp. 49-64). Cham, Switzerland: Springer.

Oller Jr, J. W. (1983). *Issues in language testing research.* Rowley, MA: Newbury.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge, England: Cambridge University Press.

Ong, J. (2014). How do planning time and task conditions affect metacognitive processes of L2 writers? *Journal of Second Language Writing, 23*(1), 17-30.

Open Door 2019 Report. (2019). Institute of International Education. Retrieved from https://www.iie.org/Research-and-Insights/Open-Doors/Data/International-Students/Places-of-Origin

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing.* Unpublished Ph.D. thesis. University of Reading, Reading, UK.

O'Sullivan, B. (Ed.). (2011). *Language testing: Theories and practices.* Oxford, England: Palgrave.

O'Sullivan, B., & Weir, C. J. (2011). *Language testing and validation. Language testing: Theory and practice.* Oxford, England: Palgrave.

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. Boston, MA: Heinle.

Oxford, R. L. (2011). Strategies for learning a second or foreign language. *Language Teaching, 44*(2), 167-180.

Oxford, R. L. (2017). *Teaching and researching language learning strategies*. New York, NY: Routledge.

Oxford, R., Cho, Y., Leung, S., & Hae-Jin, K. (2004). Effect of the presence and difficulty of task on strategy use: An exploratory study. *International Review of Applied Linguistics in Language Teaching, 42*(1), 1-47.

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (6th ed.). Sydney, NSW: Allen & Unwin.

Pallotti, G. (2019). An approach to assessing the linguistic difficulty of tasks. *Journal of the European Second Language Association, 3*(1), 58–70.

Paltridge, B., & Phakiti, A. (2010). *Continuum companion to research methods in applied linguistics.* London, England: Continuum.

Pan, Y., & Li, S. (2009). A study on the Chinese EFL learners' weak oral skill. *School Journal, 1*(1), 28-32.

Papaleontiou-Louca, E. (2003). The concept and instruction of metacognition. *Teacher Development, 7*(1), 9-30.

Papaleontiou-Louca, E. (2008). *Metacognition and theory of mind*. Newcastle, England: Cambridge Scholars.

Pei, L. (2014). Does metacognitive strategy instruction indeed improve Chinese EFL learners' reading comprehension performance and metacognitive awareness? *Journal of Language Teaching and Research, 5*(5), 1147-1152.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*(1), 26-56.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25*(2), 237-272.

Phakiti, A. (2016). Test-takers' performance appraisals, appraisal calibration, and cognitive and metacognitive strategy use. *Language Assessment Quarterly: An International Journal, 13*(2), 75-108.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language learning, 61*(4), 993-1038.

Plosky, L. (2019) Language learning strategy instructing: Recent research and future directions. In A. U. Chamot, & V. Harris (Eds.). *Learning strategy instruction in the language classroom: Issues and implementation* (pp. 3-21). Bristol, England: Multilingual Matters.

Poupore, G. (2013). Task motivation in process: A complex systems perspective. *Canadian Modern Language Review, 69*(1), 91-116.

Poupore, G. (2015). The influence of content on adult L2 Learners' task motivation: An interest theory perspective. *Canadian Journal of Applied Linguistics, 17*(2), 69-90.

Poupore, G. (2016). Measuring group work dynamics and its relation with L2 learners' task motivation and language production. *Language Teaching Research, 20*(6), 719-740.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics,15*(1), 36-57.

Powney, J., & Watts, M. (2018). *Interviewing in educational research.* New York, NY: Routledge.

Pressley, M., & Allington, R. L. (2014). *Reading instruction that works: The case for balanced teaching*. New York, NY: Guilford.

Psaltou-Joycey, A., & Gavriilidou, Z. (2018). Language Learning Strategies in Greek Primary and Secondary School Learners. In R. L., Oxford, & C. M., Amerstorfer (Eds.),

Language learning strategies and individual learner characteristics: Situating strategy use in diverse contexts (pp. 248-269). London, England: Bloomsbury.

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.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test-takers: A structural equation modelling approach. *Language Testing,15*(3), 333-379.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeming approach*. Cambridge, England: Cambridge University Press.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, England: Cambridge University Press.

Purpura, J. E. (2008). Assessing communicative language ability: Models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopaedia of language and education, language testing and assessment* (2nd ed., Vol. 7, pp. 53–68). Dordrecht, the Netherlands: Kluwer.

Purpura, J. E. (2013). Language learner strategies and styles. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th ed). Boston, MA: Heinle.

Purpura, J. E. (2016). Second and foreign language assessment. *Modern Language Journal*, 100, 190-208.

Putlack, M. A., & Link, W. (2009). *How to master skills for the TOEFL iBT Speaking Intermediate*. Beijing: Qunyan.

Qin, L. (2007). EFL teacher factors and students' affect. *US-China Education Review, 4*(3), 60-67.

Qin, T. L. (2018). *Metacognitive perspectives on learning to write in English as a foreign language (EFL) in multimedia environments at the tertiary level in China.* Unpublished Ph.D. thesis. The University of Auckland, Auckland, New Zealand.

Qin, T. L., & Zhang, L. J. (2019). English as a foreign language writers' metacognitive strategy knowledge of writing and their writing performance in multimedia environments. *Journal of Writing Research, 11*(2), 393-413.

Qin, X. (2003). *外语教学研究中的定量数据分析 [Quantitative data analysis in foreign language teaching and research*]. Wuhan, China: Huazhong University of Science & Technology Press.

Rahimi, M. (2016). Task complexity, affective factors, and pre-task planning: Effects on L2 writing production. Unpublished Ph.D. thesis. The University of Auckland, Auckland, New Zealand.

Rahimi, M., & Katal, M. (2012). Metacognitive strategies awareness and success in learning English as a foreign language: An overview. *Procedia-Social and Behavioural Sciences, 31*, 73-81.

Rahimi, M., & Zhang, L. J. (2018). Effects of task complexity and planning conditions on L2 argumentative writing production. *Discourse Processes, 55*(8), 726-742.

Rahimi, M., & Zhang, L. (2019). Writing task complexity, students' motivational beliefs, anxiety and their writing production in English as a second language. *Reading and Writing: An Interdisciplinary Journal*, *32*(3), 761-786.

Rahimirad, M., & Shams, M. R. (2014). The effect of activating metacognitive strategies on the listening performance and metacognitive awareness of EFL students. *International Journal of Listening, 28*(3), 162-176.

Rahimy, R. (2015). Examining the effect of task complexity and sequence on speaking ability of iranian EFL learners. *International Journal of Applied Linguistics & English Literature, 4*(1), 247-254.

Rashvand Semiyari, S., & Ahangari, S. (2019). The impact of task types and rating methods on Iranian EFL learners' speaking scores. *Research in English Language Pedagogy, 7*(2), 187-208.

Raudenbush, S. W., & Bryk A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Read, J. (2016). *Post-admission language assessment of university students*. Cham, Switzerland: Springer.

Reinard, J.C. (2006) *Communication research statistics.* Thousand Oaks, CA: Sage.

Révész, A. (2011). Task complexity focus on L2 constructions, and individual differences: A classroom-based study. *Modern Language Journal, 95*(4*)*, 162-181.

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics, 35*(1), 87-92.

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition, 38*(4), 703-737.

Rezazadeh, M., Tavakoli, M., & Rasekh, A. E. (2011). The role of task type in foreign language written production: Focusing on fluency, complexity, and accuracy. *International Education Studies, 4*(2), 169-176.

Richards, J. C., & Schmidt, R. W. (2013). *Longman dictionary of language teaching and applied linguistics* (4th ed.). New York, NY: Routledge.

Richards, L. (2015). *Handling qualitative data: A practical guide* (3rd ed.). Los Angeles, CA: Sage.

Robinson, P. (2001a). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Robinson, P. (2001b). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp.287-318). New York, NY: Cambridge University Press.

Robinson, P. (2003). Attention and memory during SLA. In C. Doughty, & M.H. Long (Eds.), *Handbook of Second Language Acquisition* (pp. 631-678). Oxford, England: Blackwell.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching, 43*(1), 1-32.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics in Language Teaching, 45*(3), 193-213.

Robinson, P. (2011a). Second language task complexity, the cognition hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis on language learning and performance* (pp. 3–37). Amsterdam, the Netherlands: Benjamins.

Robinson, P. (2011b). Task based language learning: A review of issues. *Language Learning, 61*(1), 1-36.

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference* (pp. 123–155). Amsterdam, the Netherlands: Benjamins.

Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching, 45*(3), 161-176.

Robinson, P., & Gilabert, R. (2020). Task-based learning: Cognitive underpinnings. In C. Chapelle (Ed.), *The concise encyclopaedia of applied linguistics* (pp. 1046-1051). Oxford, England: Blackwell.

Roderer, T., Krebs, S., Schmid, C., & Roebers, C. M. (2012). The role of executive control of attention and selective encoding for pre-schoolers' learning. *Infant and Child Development, 21*(2), 146-159.

Roebers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences, 29,* 141-149.

Roever, C., & Mcnamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics, 16*(2), 242-258.

Rong, T. S. (2011). *Amos 與研究方法 [AMOS and research methodology*]. (4th ed.). Taipei: Wu-Nan Book.

Rotenberg, A. M. (2002). A classroom research Project: The Psychological Effects of Standardized Testing on Young English Language Learners at Different Language Proficiency Levels. ERIC ED47265. Retrieved from https://eric.ed.gov/?id=ED472651.

Rubin, J. (2001). Language learner self-management. *Journal of Asian Pacific Communication, 11*(1), 25-37.

Rubin, J., Chamot, A. U., Harris, V., & Anderson, N. J. (2007). Intervening in the use of strategies. *Language Learner Strategies, 30,* 29-45.

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31-53.

Saha, M. (2014). EFL Test anxiety: Sources and supervisions. *Journal of Teaching and Teacher Education, 2*(02), 187-208.

Salataci, R. (2002). Possible effects of strategy instruction on L1 and L2 reading. *Reading in a Foreign Language, 14*(1), 1-17.

Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Los Angeles, CA: Sage.

Salimi, A., & Dadashpour, S. (2012). Task complexity and SL development: Does task complexity matter? *Procedia-Social and Behavioural Sciences*, 46, 726-735.

Salkind, N. J. (2010). *Encyclopaedia of research design*. Los Angeles, CA: Sage.

Samejon, K. (2015). Extrinsic motivation factors in learning English as second language. *Kaalam, 1*(1), 35-53.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Basingstoke, England: Palgrave.

Sasayama, S. (2015). *Validating the assumed relationship between task design, cognitive complexity, and second language task performance.* Unpublished Ph.D. thesis. Georgetown University, Washington, DC, the USA.

Sasayama, S. (2016). Is a 'complex 'task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal, 100*(1), 231-254.

Saraswati, D. (2017). The relationship between reading strategy use and EFL test performance. *Journal of NELTA, 22*(1-2), 38-52.

Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics, 40*(6), 894-916.

Schellings, G., & Hout-Wolters, B. (2011). Measuring strategy use with self-report instruments: Theoretical and empirical considerations. *Metacognition and Learning, 6*(2), 83-90.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology, 31*(1), 64-96.

Scott, J. (2014). *Factor analysis* (4th ed.) Oxford, England: Oxford University Press.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(1-4), 209-232.

Seong, Y. (2014). Strategic competence and L2 speaking assessment. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 14*(1), 13-24.

Setiyadi, A., Mahpul, Sukirlan, M., & Rahman, B. (2016). Language motivation, metacognitive strategies and language performance: A cause and effect correlation. *International Journal of Applied Linguistics and English Literature, 5*(7), 40-47.

Shen, Y., & Zhang L. J. (2006). Investigating the relationship between metacognition and EFL listening of postgraduate non-English majors: A multiple regression analysis. *Foreign Language Research, 3*(157), 45-52.

Shih, H., & Huang, S. (2018). EFL learners' metacognitive strategy use in reading tests. *English Teaching & Learning, 42*(2), 117-130.

Shih, H. C. J., & Huang, S. H. C. (2020). College students' metacognitive strategy use in an EFL flipped classroom. *Computer Assisted Language Learning, 33*(7), 755-784.

Shohamy, E., Or, I. G., & May, S. (2017). *Language testing and assessment* (3rd ed.). Cham, Switzerland: Springer.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied linguistics, 17*(1), 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.

Skehan. P (2011). *Researching tasks: Performance, assessment and pedagogy*. Shanghai: Shanghai Foreign Language Education Press.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. Jackson, & Y. Yilmaz (Eds.) *Cognitive individual differences in second language processing and acquisition* (pp.7-40). Amsterdam, the Netherland: Benjamins.

Skehan, P. (2018). *Second language task-based performance: Theory, research, assessment.* New York, NY: Routledge.

Ślęzak-Świat, A. (2008). *Components of strategic competence in advanced foreign language users*. Unpublished Ph.D. thesis. University of Silesia, Katowice, Poland.

Snijders, T. A. B. (2012). In R. J. Bosker (Ed.), *Multilevel analysis: An introduction to basic and advanced multilevel modelling* (2nd ed.). Los Angeles, CA: Sage.

Song, X. (2005). Language learner strategy use and English proficiency on the Michigan English Language Assessment Battery. *Spain Fellow Working Papers in Second or Foreign Language Assessment, 3,* 1-26.

Song, X., & Cheng, L. (2006). Language learner strategy use and test performance of Chinese learners of English. *Language Assessment Quarterly: An International Journal, 3*(3), 243-266.

Sperling, R. A., Richmond, A. S., Ramsay, C. M., & Klapp, M. (2012). The measurement and predictive ability of metacognition in middle school learners. *The Journal of Educational Research, 105*(1), 1-7.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge, England: Cambridge University Press.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence*. New York, NY: Viking.

Sun, D. (2014). From Communicative Competence to Interactional Competence: A New Outlook to the Teaching of Spoken English. *Journal of Language Teaching & Research*, *5*(5), 1062-1070.

Sun, P. P. (2016). *Chinese as a second language learners' speech competence and speech performance in classroom contexts: Cognitive, affective, and socio-cultural perspectives.* Unpublished Ph.D. thesis. The University of Auckland, Auckland, New Zealand.

Sun, P., Zhang, L., & Gray, S. (2016). Development and validation of the speaking strategy inventory for learners of Chinese (SSILC) as a second/foreign language. *The Asia-Pacific Education Researcher, 25*(4), 593-604.

Swain, M., Huang, L., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT™ (sstibt): Test-takers' reported strategic behaviours* (Research Report No. RR-09-02). Princeton, NJ: Educational Testing Service.

Tajeddin, Z., & Bahador, H. (2012). Pair grouping and resource-dispersing variables of cognitive task complexity: Effects on L2 output. *Iranian Journal of Applied Linguistics (IJAL), 15*(1), 123-149.

Takeuchi, O. (2020) Language learning strategies: Insights from the past and directions for the future. In X. Gao (Ed.), *Second handbook of English language teaching* (pp. 684-699). Cham, Switzerland: Springer.

Tarone, E. (2005). Speaking in a second language. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 485–502). Mahwah, NJ: Erlbaum.

Tarricone, P. (2011). *The taxonomy of metacognition*. Hove, England: Psychology.

Tavakoli, P. (2009). Investigating task difficulty: learners' and teachers' perceptions. *International Journal of Applied Linguistics, 19*(1), 1-25.

Taylor, L. B. (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge, England: Cambridge University Press.

Teng, L. S. (2016). *Fostering strategic second-language writers: A study of Chinese English -as- a- foreign language (EFL) writers' self-regulated learning strategies.* Unpublished Ph.D. thesis. The University of Auckland, Auckland, New Zealand.

Teng, L. S., Sun, P. P., & Xu, L. (2018). Conceptualizing writing self-Efficacy in English as a foreign language context: Scale validation through structural equation modelling. *TESOL Quarterly, 52*(4), 911-942.

Teng, L. S., & Zhang, L. J. (2016). A Questionnaire-based validation of multidimensional models of self-regulated learning strategies. *Modern Language Journal, 100*(3), 674-701.

Teng, L. S., & Zhang, L. J. (2020). Empowering learners in the second/foreign language classroom: Can self-regulated learning strategies-based writing instruction make a difference? *Journal of Second Language Writing, 48*, 1-12.

Tian, J. (2012). 浅议雅思托福对大学英语教育的影响 [The influence of IELTS and TOEFL on China's EFL teaching in universities]. *Social Science Review,27* (6), 173-174.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119.

Tran, T. H. (2012). *Second language assessment for classroom teachers.* Paper presented at MIDTESOL 2012, Ames, Iowa, ERIC ED545800. retrieved from https://eric.ed.gov/?id=ED545800

Tsagari, D., & Banerjee, J. (2016). *Handbook of second language assessment.* Boston, MA: De Gruyter Mouton.

Tubbs, J. M. (2016). Guidebook for ESL teachers of Chinese students. Master's Projects and Capstones (326). Retrieved from https://repository.usfca.edu/capstone/326.

Upton, G., & Cook, I. (2014*). Inter-rater reliability* (3rd ed.). Oxford, England: Oxford University Press.

Vandergrift, L., Goh, C. C., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning, 56*(3), 431-462.

Veenman, M. (2011). Alternative assessment of strategy use with self-report instruments: A discussion. *Metacognition and Learning, 6*(2), 205-211.

Veenman, M., Hout-Wolters, B., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning,1*(1), 3-14.

Veenman, M. V., & van Cleef, D. (2019). Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods. *ZDM International Journal on Mathematics Education, 51*(4), 691–701.

Verma, J. P. (2015). *Repeated measures design for empirical researchers*. Hoboken, NJ: Wiley-Blackwell.

Vrettou, A. (2009). Language learning strategy employment of EFL Greek-speaking learners in junior high school. *Journal of Applied Linguistics, 25*, 85-106.

Vrettou, A. (2011). *Patterns of language learning strategy use by Greek-speaking young learners of English.* Unpublished Ph.D. Thesis. Aristotle University of Thessaloniki, Thessaloniki, Greece.

Walford, G. (2001). *Doing qualitative educational research: A personal guide to the research process.* London, England: Continuum.

Wang, D., Lai, H., & Leslie, M. (2015). Chinese English learners' strategic competence. *Journal of Psycholinguistic Research,44*(6), 701-714.

Wang, H., & Yu, G. (2018). Test-Takers' Cognitive Processes during a Listen-To-Summarize Cloze Task. *International Journal of Listening, 35(*1), 1-28.

Wang, L., & Zhang, L. J. (2019). Peter Skehan's influence in research on task difficulty in Z. E. Wen, & M. J. Ahmadian (Eds). *Researching L2 task performance and pedagogy: In honour of Peter Skehan* (pp.183–198). Philadelphia, PA: Benjamins.

Wang, R., & Liu, Y. (2018*).* Literature review of syntactic priming experiment methods and bilingual speech production models. *Advances in Social Science, Education and Humanities Research, 283,* 394-398.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Hampshire, England: Palgrave.

Weir, C. J., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in speaking tasks: An intra-task perspective. *IELTS Research Report No.6* (pp. 119–160). British Council and IDP Australia.

Wei, R., & Su, J. (2015). Surveying the English language across China. *World Englishes, 34*(2), 175-189.

Wen, Y. (2016). 大学英语教学中情感因素的影响——理论与实证研究. [The role of affective factors in Chinese EFL teaching in universities: An empirical study]. *Overseas English,* 23, 84-86.

Wenden, A. (1987). Metacognition: An expanded view on the cognitive abilities of L2 learners. *Language Learning*, *37*(4), 573-597.

Weng, F, X. (2009). 阶层线性模型的原理与应用. [*The theory and applications of multilevel modelling*]. Beijing, China: China Light Industry Press.

Weng, F. X., & Qiu, H. Z. (2014). 多层次模式方法论：阶层性模式的关键问题及试解. [*Methodology of multilevel modeming: Key questions and solutions on multilevel modelling*] (3rd ed.). Beijing, China: Press of Economic Management.

Wilson, A. (2013). To be specific: *Using professional development with subject-specific and generic literacy components to raise secondary students' achievement in English, mathematics and science.* Unpublished Ph.D. thesis. The University of Auckland, Auckland, New Zealand.

Wischgoll, A. (2016). Combined training of one cognitive and one metacognitive strategy improves academic writing skills. *Frontiers in Psychology*, *7*(187), 1-13.

Woods. M., (2009). *A spiritual journey through poetry with Marion Woods*. Bloominton, IN: Authourhouse.

Xie, K., & Bradshaw, A. C. (2008). Using question prompts to support ill-structured problem solving in online peer collaborations. *International Journal of Technology in Teaching and Learning, 4*(2), 148-165.

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modelling. *Language Testing, 30*(1), 49-70.

Xu, J. (2016). The relationship between the use of speaking strategies and performance on IELTS speaking test: A study on Chinese college students. *International Journal for 21st Century Education, 3*(2), 69-96.

Xu, T. S., Zhang, L. J., & Gaffney, J. (2021). Examining the relative effectiveness of task complexity and cognitive demands on students' writing in a second language. *Studies in Second Language Acquisition.* Accepted

Xu, X. (2011). The relationship between language learning motivation and the choice of language learning strategies among Chinese graduates. *International Journal of English Linguistics, 1*(2), 203-212.

Yang, H. C. (2014). Toward a model of strategies and summary writing performance. *Language Assessment Quarterly: An International Journal, 11*(4), 403-431.

Yang, H., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly, 46*(1), 80-103.

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education, 23*(1), 31-48.

Yahya, M. Y. (2019). *Improving speaking performance and L2 motivation through task-based language teaching on Malaysian undergraduate students.* Unpublished Ph.D. thesis. The University of Reading, Reading, UK.

Yi, J. I. (2012). *Comparing strategic processes in the iBT speaking test and in the academic classroom*. Unpublished Ph.D. thesis. The University of Leicester, Leicester, UK.

Yi-Ching Pan, & Yo In'nami. (2015). Relationships between strategy use, listening proficiency level, task type, and scores in an L2 listening test. *Canadian Journal of Applied Linguistics, 18*(2), 45-77.

Yin, Z. (2013). Infer the meaning of unknown words by sheer guess or by clues? An exploration on the clue use in Chinese EFL learner's lexical inferencing. *English Language Teaching, 6*(11), 29-38.

Youn, S. J., & Bi, N. Z. (2019). Investigating test-takers' strategy use in task-based L2 pragmatic speaking assessment. *Intercultural Pragmatics, 16*(2), 185-218.

Young, D. S. (2017). *Handbook of regression methods.* Boca Raton: CRC.

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., & Fang, L. (2017). *Preparing for the speaking tasks of the TOEFL iBT® test: An investigation of the journeys of Chinese test-takers* (Research Report No. RR-17–01). Princeton, NJ: Educational Testing Service.

Yu, G., & Zhang, J. (2017). Computer-based English language testing in China: Present and future. *Language Assessment Quarterly: An International Journal, 14*(2), 177-188.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologist oral production. *Applied Linguistics, 24*(1), 1-27.

Zare, P. (2012). Language learning strategies among EFL/ESL learners: A review of literature. *International Journal of Humanities and Social Science, 2*(5), 162-169.

Zeidner, M. (1998). In I., Ebrary (Ed.), *Test anxiety: The state of the art*. New York, NY: Kluwer.

Zhang, D., & Goh, C. C. M. (2006). Strategy knowledge and perceived strategy use: Singaporean students' awareness of listening and speaking strategies. *Language Awareness, 15*(3), 199-119.

Zhang, D., & Zhang, L. J. (2019). Metacognition and self-regulated learning (SRL) in second/foreign language teaching. In X. Gao (Ed.), *Second handbook of English language teaching* (pp. 883-897). Cham, Switzerland: Springer.

Zhan, J. (2006). 听说并重，口语领先——略论中国英语口语教学中的三个对立统一关系[Equal attention to listening and speaking with special focus on speaking: A study of the

Chinese EFL classroom instructions on speaking]. *Technology Enhanced Foreign Language Education, 5*, 55-59.

Zhang, L. (2014). A structural equation modelling approach to investigating test-takers' strategy use and reading test performance. *Asian EFL Journal, 16*(1), 152-188.

Zhang, L. (2016). Chinese college test-takers' individual differences and reading test performance: A structural equation modelling approach. *Perceptual and Motor Skills, 122*(3), 725-741.

Zhang, L. (2017). *Metacognitive and cognitive strategy use in reading comprehension: A structural equation modelling approach*. Singapore: Springer.

Zhang, L., Aryadoust, V., & Zhang, L. J. (2014). Development and validation of the test-takers' metacognitive awareness reading questionnaire (TMARQ). *The Asia-Pacific Education Researcher, 23*(1), 37–51.

Zhang, L., Goh, C. C., & Kunnan, A. J. (2014). Analysis of test-takers' metacognitive and cognitive strategy use and EFL reading test performance: A multi-sample SEM approach. *Language Assessment Quarterly: An International Journal*, *11*(1), 76-102.

Zhang, L., & Zhang, L. J. (2013). Relationships between Chinese college test-takers' strategy use and EFL reading test performance: A structural equation modelling approach. *RELC Journal*, *44*(1), 35-57.

Zhang, L. J (1999). *Metacognition, cognition and L2 reading: A study of Chinese university EFL readers' metacognitive knowledge and strategy deployment.* Unpublished Ph.D. thesis. Nanyang Technological University, Singapore.

Zhang, L. J. (2010). A dynamic metacognitive systems account of Chinese university students' knowledge about EFL reading. *TESOL Quarterly, 44*(2), 320-353.

Zhang, L. J., Aryadoust, V., Zhang, D. (2016) Taking stock of the effects of strategies-based instruction on writing in Chinese and English in Singapore primary schools. In R. E. Silver, & W. Bokhorst-Heng (Eds.), *Quadrilingual education in Singapore: Pedagogical innovation in language education (*pp. 103–126). New York, NY: Springer.

Zhang, L. J., & Qin, T. L. (2018). Validating a questionnaire on EFL writers' metacognitive awareness of writing strategies in multimedia environments. In A. Haukås, C. Bjørke,

& M. Dypedahl (Eds.), *Metacognition in language learning and teaching* (pp. 157–179). London, England: Routledge.

Zhang, L. J., Thomas, N., & Qin, T. L. (2019) Language learning strategy research in System: Looking back and looking forward. *System, 84*, 87-93.

Zhang, L. J., & Xiao, Y. (2006). Language learning strategies, motivation and EFL proficiency: A study of Chinese tertiary-level non-English majors. *Asian Englishes: An International Journal of the Sociolinguistics in Asia/Pacifc*, *9*(2), 20-47.

Zhang, L. J., & Zhang, D. (2018). Metacognition in TESOL: Theory and practice. In J. I. Liontas, A. Shehadeh (Eds.), *The TESOL encyclopaedia of English language teaching* (Vol. II, pp. 682-792). Hoboken, NJ: Wiley-Blackwell.

Zhang, Y., & Huang, L. (2019). 改革开放 40 年中国高校国际化的成就与挑战 [Achievements and challenges related to the globalisation of Chinese universities 40 years after China's reform and opening up]. *北京理工大学学报* [*Journal of Beijing University of Technology]*, *2019*(1), 27-36.

Zhao, W. L. (2017). Issues on China's ELT assessment. *Education for Chinese Adults*, 18, 147-151.

Zhao, Z. (2013). An overview of models of speaking performance and its implications for the development of procedural framework for diagnostic speaking tests. *International Education Studies*, *6*(3), 66.

Zhou, L. (2020). 浅谈高中国际部英语口语教学的突破[On how to make a breakthrough in teaching Chinese high school EFL learners speaking skill in international schools]. 《课程教育研究》*[Curriculum Education Research]*, 8, 112-113.

Zhu, X., Liao, X., & Cheong, C. M. (2019). Strategy use in oral communication with competent synthesis and complex interaction. *Journal of Psycholinguistic Research, 48*(5), 1163-1183.

Zimmerman, B. J., & Schunk, D. H. (2011). *Handbook of self-regulation of learning and performance*. New York, NY: Routledge.