Investigating the Influence of Age and Language Aptitude on the Implicit and Explicit Knowledge of ESL Learners

Lan Wei

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Applied Linguistics, The University of Auckland, 2019.

ABSTRACT

Learning age and language aptitude have been identified as two of the most important individual difference factors that influence the long-term learning outcomes of second language (L2) learners. Language aptitude, as comprised of a set of cognitive abilities, mediates the negative influence of learning age, and thus helps to make up for a late start in learning an L2 (DeKeyser, 2000). However, in terms of understanding the nature of the interaction of learning age and language aptitude, a consensus has not yet been made.

One theoretical postulation that explains the joint effect of learning age and language aptitude makes references to the different learning systems, that is, implicit and explicit learning systems that are involved in L2 learning (Bley-Vroman, 1988). It is hypothesised that language aptitude plays a compensatory role by facilitating the explicit learning system of learners, so that the late learners whose implicit learning system is less active would have a chance to reach a relatively high level of achievement in their L2.

A number of studies have been carried out empirically to examine if language aptitude indeed interacts with learning age as postulated above. However, the research findings have been mixed. Some scholars have found that language aptitude is connected with the explicit learning system only (DeKeyser, 2000; DeKeyser, Alfi-Shabtay & Ravid, 2010; Granena & Long, 2013a), while others concluded that language aptitude also plays a role in implicit learning (Abrahamsson & Hylstenstam, 2008; Granena, 2014). These mixed findings call for further research.

This study set out to investigate the extent to which age and language aptitude influence long-term L2 learning outcomes, which are explained in terms of a distinction between implicit and explicit knowledge. A battery of tests was employed and validated as measures of implicit L2 knowledge, explicit L2 knowledge and language aptitude. A total number of 20 native speakers (NS) of English and 86 non-native speakers (NNS) whose native language were Mandarin took part in this

research. At the time of data collection, all of the NNS had been living in the L2 country (New Zealand) for more than 8 years and had been using English as a medium for study and work throughout this period.

Results show that learning age is correlated with implicit L2 knowledge, which shows that age indeed influences the implicit learning system of the NNS. Language aptitude is connected with both implicit and explicit knowledge, showing that language aptitude is probably pertinent to both implicit and explicit learning. These results indicate that a young learning age and a high level of language aptitude are both necessary conditions for developing a high level of implicit L2 knowledge. In other words, a high level of language aptitude is not only advantageous for adult learners, but also for early/child learners of L2.

To my parents

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help of many individuals. I was most fortunate to have had the opportunity to work with many people, who inspired me and contributed to my work in various ways.

As I am writing this, I feel that my gratitude to my main supervisor, Dr. Rosemary Erlam, is beyond my ability to put it into words in English, just like the many times when I felt that there was a gap between my thinking and writing of my research findings. My utmost gratitude goes to Rosemary, for her tremendous support and guidance throughout the years of the study. Her expertise, insights, encouragement, and generous help have made my PhD journey exceptionally enjoyable. I would not have been able to complete this study without her. To me, Rosemary is the most caring and inspiring person I know, and she always has the magic power to make me believe in my abilities whenever I experience difficulties.

I would like to extend my sincere thanks to my co-supervisor, Dr. Shaofeng Li, for his valuable help and advice. I was fortunate to have had the opportunity to work as a teaching assistant and a research assistant for Dr. Li, and I have greatly benefited from this experience.

I would also like to say thank you to my friends in Evangelical Formosan Church of Central Auckland and St George's Church, Epsom. I especially thank Wenju Lao, Frances Chen, Blandon Leung, Hilary Sussex, and Josh Jones, for helping me with participant recruitment. Without their friendship and support, I would not have been able to have had the chance to meet many of my participants. I am grateful to all my participants, for their willingness to participate in this study and to complete a series of demanding linguistic and language aptitude tests.

I would like to take this opportunity to thank my parents Prof. Qifeng Wei and Prof. Xiulian Ren, for their unfailing love, understanding and support. My parents have been my academic role models, and they have motivated me to pursue a doctoral degree overseas.

Last but not least, a very special thank you goes to my husband Di Wu, who encourages and takes care of me, while working tirelessly on his own PhD research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AccuracyG	Accuracy/sum scores of the grammatical sentences
AccuracyUG	Accuracy/sum scores of the ungrammatical sentences
AIC	Akaike information criterion
ANOVA	Analysis of variance
AO	Age of onset
AoA	Age of arrival
AoE	Age of exposure
AoT	Age of testing
BIC	Bayesian information criterion
CANAL-F	Cognitive Ability for Novelty in Acquisition of Language (Foreign)
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CPH	Critical period hypothesis
DLAB	Defense language aptitude battery
DP model	Declarative procedural model
EFA	Exploratory factor analysis
EFL	English as a foreign language
EI	Elicited imitation test
EK	Explicit knowledge
EM	Expectation-maximization
ERP	Event-related potential
ESL	English as a second language
fMRI	Functional magnetic resonance imaging
GJT	Grammaticality judgement test
GSI	Grammaticality sensitivity index
IK	Implicit knowledge
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
L1	First language

L2	Second language
LoR	Length of residence
MAD	Median absolute deviation
MANOVA	Multivariate analysis of variance
MCAR	Missing completely at random
MKT	Metalinguistic knowledge test
ML	Maximum likelihood
MLAT	Modern language aptitude test
MLM	Maximum likelihood with robust standard errors and χ^2
MTG	Monitor times of the grammatical sentences
MTMM	Multi-trait multi-model
MTUG	Monitor times of the ungrammatical sentences
NS	Native speakers
NNS	Non-native speakers
NonWord	Nonword repetition test
PCA	Principal component analysis
PET	Positron emission tomography
PLAB	Pimsleur language aptitude battery
RMSEA	Root mean square error of approximation
RT	Response time
SPH	Sensitive period hypothesis
SRMR	Standardised mean squared residual
SSpan	Listening sentence span test
TGJT	Timed grammaticality judgement test
TLI	Tucker-Lewis index
UGJT	Untimed grammaticality judgement test
WordM	Word monitoring test

CHAPTER ONE

INTRODUCTION

This chapter focuses on issues underpinning this research. It will firstly present a brief introduction of the extant research on age and language aptitude. Then, the learning contexts of the participants are introduced, followed by the personal incentives of the researcher and the rationale for conducting the present study. Finally, several conceptual distinctions are made, and the outline of the thesis is presented.

1.1 Research background

A common observation that people repeatedly make regarding the effect of age on language acquisition is that children are better at learning languages than adults. This general observation is probably based on the fact that, firstly, in relation to the acquisition of one's first language (L1), children reach perfect mastery of whatever language they are exposed to; and that secondly, children generally reach a higher level of proficiency in acquiring a second language (L2) in comparison to adults. When the outcomes of first language acquisition and second language acquisition are compared, second language acquisition is characterised by considerable individual variation in the level of L2 proficiency attained whereas first language acquisition normally leads to uniform success. This stark difference in terms of language acquisition outcomes, has therefore attracted the attention of researchers for decades and both theoretical inquiries and empirical studies have been conducted in order to understand why this is the case.

Much of previous research has been conducted within the theoretical framework of a postulated *critical period* of language acquisition (Johnson & Newport, 1989; Lenneberg, 1967; Long, 1990), which predicts that nativelike attainment in an L2 will not be possible if second language acquisition starts after

this finite period. A considerable amount of evidence has been provided in support of this critical period hypothesis (CPH), showing that children who start to acquire the L2 at an age within the critical period (usually before puberty) are able to become nativelike (e.g. Birdsong, 1999). However, there are also instances where some adult learners become nativelike, and they even perform as well as native speakers in some demanding psycholinguistic experiments (e.g. Abrahamsson & Hyltenstam, 2008). Therefore, these learners are taken as examples to argue against the CPH. Indeed, the age issue has been one of the most disputed topics in the field of second language acquisition (Hyltenstam & Abrahamsson, 2000), and a widely accepted explanation of the effect of age is yet to be made.

In addition to age, another factor that has been extensively researched is language aptitude. Just as age is recognised as having a role in second language acquisition, language aptitude also has been regarded as one of the most robust factors that influence the outcomes of second language acquisition. A number of recent studies have suggested a possible interplay between age and aptitude, in the sense that high language aptitude could compensate for the negative influence of age (Abrahamsson & Hyltenstam, 2008; DeKeyser, Alfi-Shabtay, & Ravid, 2010; Granena & Long, 2013a). At a theoretical level, this possible interplay has been explained with reference to the fundamental difference hypothesis (Bley-Vroman, 1988), which claims the learning mechanisms for children and adults (including post-adolescents) in second language acquisition are different. Following this hypothesis, DeKeyser (2000) predicts that aptitude, especially verbal analytical ability, compensates for decreasing learning ability as learners grow older. He further argues that high levels of aptitude make it possible for adult learners to reach a nativelike level of L2 proficiency. To date, there are only a handful of studies investigating this possible interplay between age and aptitude (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; DeKeyser et al., 2010; Granena & Long, 2013a; Granena, 2014). Therefore, the present study aims to make a contribution to this research area by empirically testing implicit and explicit knowledge independently

and then examining the relevance of age and language aptitude in relation to these two types of knowledge.

1.2 English learning contexts in China and New Zealand

In this study, 86 native speakers of Mandarin Chinese whose second language is English are involved (see Section 3.1 for more details). These learners generally have the experience of learning English in two different contexts, namely China and New Zealand. In this section, the learning contexts are briefly introduced.

The English learning context in China is a traditional instructed context in which English is mainly taught as a foreign language in classrooms, and the purpose is largely for examinations. The opportunity to learn a foreign language has held an important place in the Chinese education system, and English has been the most popular language that students choose to learn as a subject. English is introduced to students as a compulsory course in primary school and is taught throughout the years in secondary school and university in most cases. However, the use of English is generally restricted to language classrooms, and activities that use English as a medium for communication are far and few between. Moreover, Chinese students tend to put a lot of effort into memorising English grammar rules, in order to get high scores in the English test of the National Higher Education Entrance Examination. Since the exam does not include a section testing speaking, students' ability to use English in oral communication has been an aspect that generally receives little attention in English classrooms. In other words, the kind of learning happens in Chinese English classrooms is explicit in nature (see Section 2.2.1), as it is the common practice for teachers to explicitly teach vocabulary and grammar and then to ask the students to memorise this knowledge. As a result, Chinese students tend to have considerable explicit knowledge (see Section 2.2.2), and they would mainly use their explicit knowledge to complete the English test of the National Higher Education Entrance Examination.

In comparison, learning English in a naturalistic context, such as the context of this study in New Zealand, indicates the exposure to a larger amount of linguistic input and also the use of English for communicative purposes. In such a learning context, it is more likely that implicit learning occurs, provided that learners' attention is not explicitly drawn to aspects such as grammar (see Section 2.2.1). As a result of implicit learning and also consistent English use in oral communication, some implicit knowledge could be developed (see Section 2.2.2). However, for some Chinese learners of English, this may not be the case because there is a large Chinese population living in New Zealand. The latest census data shows that there were 171,411 Chinese people living in New Zealand, making up 4.3% of the total population in 2013 (StatsNZ, 2015). Of these Chinese residents, 69% were living in Auckland, the city in which this study was conducted. There are grocery stores, real estate agencies, travel agencies and other businesses owned and operated by Chinese people and more and more businesses are providing Chinese language services nowadays. This social environment makes it relatively easy for Chinese immigrants to live in Auckland and use Chinese in their daily lives. Nevertheless, English must be used for other purposes, for example, studying in schools or universities or undertaking professional work (unless Chinese is required as the working language). This context would probably influence the learning outcomes of Chinese learners, as they may not be compelled to use English outside their studying or working environments. In summary, the extent to which Chinese learners of English have the opportunity to acquire English through exposure to the language in a naturalistic context in New Zealand could vary, depending on their daily lives, working environments and the education they receive.

1.3 Personal incentives

The present study is motivated to some extent by the personal experience of the researcher. Like the majority of the participants in this study, the researcher is a

Chinese second language learner of English, who has the experience of both learning English in an instructed context (i.e. China) and in a naturalistic context (i.e. New Zealand). However, it appears that albeit being very proficient, she feels that she is far from being nativelike in English. This feeling drives her to reflect on her English learning experience.

The researcher received classroom English instruction from primary school to secondary school because English is a compulsory subject in the Chinese education system. She also had some additional experience in learning English outside the classroom. To start with, the researcher was exposed to English at a relatively young age. She was sent to study English at the age of 7 because her mother was very much influenced by the so-called "the earlier, the better" belief about English learning, which was prevalent in China at that time. After that, she started learning English as a subject in primary school at the age of 10. In the meantime, she attended extra-curricular English lessons once a week for an hour and studied with English teachers who were native English speakers. Perhaps due to this experience of learning English with native English speakers, the researcher had never found English difficult at school, but she noticed that this subject had obviously caused her classmates difficulty. When the researcher started attending secondary school, she received more English instruction, but she no longer attended extra-curricular English lessons. English was even more emphasised as she began to prepare for the National Higher Education Entrance Examination, but at that time lessons were mainly focused on exam preparation, which included multiple choice questions, cloze exercises, reading comprehension and writing exercises, all for the purpose of getting high scores in the exam.

After graduation from secondary school, the researcher studied for a Bachelor of Arts degree in a top Chinese university where she took English as the major. During her undergraduate study, she attended courses that further trained her in the four skills, that is, listening, speaking, reading and writing. The classroom instruction that she received in university was not much different from the

instruction that she had received in secondary school, but she had many more opportunities to express her ideas using English in university. Following her undergraduate study, the researcher continued to study for a Master of Arts degree in Applied Linguistics at the same university.

When the researcher moved to New Zealand to study for her PhD in Applied Linguistics, she began to realise that although she was very proficient in English, there was a big gap between her language competence and that of native speakers. She found that in order to avoid making grammatical mistakes in oral communication, she had to constantly remind herself about the grammar rules that she had learnt. The correct use of count and non-count nouns and subject verb agreement were areas that she noticed were particularly difficult. She made mistakes where singular collective nouns or non-count nouns were used in plural forms. The corresponding grammar rules are undoubtedly fundamental because they were taught at an early stage in her English classes, but it appears that the researcher lacks accuracy and automaticity (i.e. implicit knowledge) in applying these rules in spontaneous communication. In other words, it seems that the researcher has considerable grammatical knowledge of English (i.e. explicit knowledge), but still makes some grammatical mistakes in oral communication.

This phenomenon seems to be a common characteristic of many other Chinese learners of English as the researcher noticed that the Chinese people she met in New Zealand tended to have the same problems. This observation made the researcher wonder why Chinese speakers of English make grammatical mistakes in oral communication even if they have learnt English for many years in different contexts. She also wondered whether or not it is possible for them to speak English like native speakers, that is, making minimal grammatical mistakes without constantly reminding themselves about these grammatical structures in spontaneous speech. These personal experiences drove the researcher to conduct a study with an aim of understanding the long-term learning outcomes of Chinese learners of English and exploring the factors that influence the learning outcomes.

1.4 Conceptual issues: some basic distinctions

Naturalistic vs. instructed learning

In the present study, a distinction is made between naturalistic and instructed learning of an L2. In the case of learning English as the L2, naturalistic learning usually occurs in English as a second language (ESL) contexts, whereas instructed learning normally refers to the type of learning that takes place in English as a foreign language (EFL) contexts. It should be noted, however, that instructed learning can occur in an ESL context as well, for example, in language schools where English is taught to students with an emphasis on grammatical correctness. To put it another way, naturalistic learning can be characterized as "learning through immersion in the L2 environment" (Muñoz, 2008, p. 578), whereas instructed learning refers to formal learning in L2 classrooms.

The distinction between naturalistic and instructed L2 learning is important because the quantity and quality of input in these contexts differ significantly, and findings from ESL contexts are not fully applicable to EFL contexts (DeKeyser & Larson-Hall., 2005). In EFL contexts, L2 exposure is normally restricted to classroom settings, and the L2 is not used as a medium for communication after class. Also, since teachers are the main source of the target language English, there are large variations in terms of the quality (teachers vary in oral fluency and general proficiency) and quantity (the target language is not always used in teaching) of linguistic input (Muñoz, 2008). In contrast, in a naturalistic environment, learners are given the opportunity to learn through the exposure to abundant language resources, and it is usually under this circumstance that near-native learners are found.

In addition, the distinction between naturalistic and instructed learning denotes a difference in the type of learning that is involved. In naturalistic learning contexts where learners are immersed in the L2 environment, it is more likely that implicit learning is taking place because learners acquire the L2 through mere exposure. In comparison, explicit learning is more likely to occur when learners are learning in

instructed learning contexts because they are probably instructed in a way which involves grammar, vocabulary, and other aspects of an L2 being taught explicitly. Any reference to naturalistic and instructed L2 learning therefore, implies that it is likely that there is a difference in the type of learning involved.

Acquisition vs. learning

The distinction between acquisition and learning (Krashen, 1982) is widely accepted because it describes two independent and distinct ways that competence in an additional language can be developed. Acquisition refers to an implicit process, which generally occurs in naturalistic learning contexts. In comparison, in instructed contexts it is commonly held that learning rather than acquisition takes place. To put it another way, the acquisition of an L2 is a subconscious process that resembles the acquisition of one's L1, whereas the learning of an L2 involves conscious learning of grammar, vocabulary and other linguistic aspects. Krashen (1982) points out that adults have the ability of both acquiring and learning an L2. Just as children acquire their L1 through exposure to that language, adults have the ability to acquire languages in a similar fashion. However, the ability of adults might be very limited in comparison with children.

The present study explores this distinction because the participants had the experience of learning their L2 in an instructed context (i.e. China) and acquiring the L2 in a naturalistic context (i.e. New Zealand). By making a distinction between acquisition and learning, inferences are made with regard to the involvement of different kinds of learning processes (i.e. implicit and explicit learning) in the acquisition of an L2. Other than these, the broader term acquisition, as in second language acquisition, is used to refer to general language acquisition.

Rate of learning vs. ultimate attainment

Another distinction of central importance is that between the rate of learning and ultimate attainment. This distinction, put forward by Krashen, Long & Scarcella

(1979), demonstrates two kinds of advantages. First, there is a rate advantage for older children, adolescents, and adults over younger children at early stages of language learning. Supporting evidence for this mainly comes from research in instructed settings, in which older learners are found to progress faster than younger learners given the same amount of instruction, which tends to be explicit in nature (e.g. Lecumberri & Gallardo, 2003; Muñoz, 2006). This rate advantage is probably a result of the superior cognitive ability of older learners and adults, that is, their explicit learning ability. In comparison, there is a long-term ultimate attainment advantage for younger learners over older learners. The large numbers of studies of immigrants in naturalistic settings provide evidence of this kind, showing that the earlier one starts to be exposed to the L2 environment, the more possible it is that one reaches a higher level of L2 proficiency. This long-term advantage in L2 attainment is likely associated with the better implicit learning ability of younger learners in the sense that they are more able to acquire the L2 by being immersed in naturalistic learning contexts than older learners and adults. In other words, given the same naturalistic learning context, the advantageous implicit learning ability of younger learners would eventually lead to better attainment in an L2 in comparison with older learners whose implicit learning systems are less active.

By distinguishing between rate of learning and ultimate attainment, researchers are able to examine the human capacity of second language acquisition. Only those studies that investigate ultimate attainment can speak of human learning capacity, whereas studies that examine the rate of learning cannot. However, it should be noted that ultimate attainment does not necessarily mean nativelikeness; rather, it suggests the end state of learning. It refers to the final product of second language acquisition regardless of the level of proficiency in the L2, and it includes any and all second language acquisition end points, up to and including nativelikeness (Birdsong, 2009).

The present study is concerned with ultimate attainment because the primary aim is to examine the extent to which age and language aptitude influence second

language acquisition. In other words, the present study sets out to investigate whether the learning potential of second language learners is influenced by age, aptitude, or a combination of these two factors. This research orientation obviously calls for attention to variation in the long-term outcomes of second language acquisition, rather than variation in the short term learning rate (Long, 1990; 2005). The term ultimate attainment is mainly used in the literature review of this thesis, as this term has been used previously. In discussing the research findings of the present study, ultimate attainment is referred to as long-term L2 outcomes to avoid the connotation of making an absolute judgement about whether what was measured in the present study was in fact the ultimate level of proficiency that the L2 learners in this study could ever achieve.

1.5 Rationale and possible implications

Inspired by research findings to date, the present study aims to discover the roles of age and aptitude in relation to L2 learners' implicit and explicit L2 knowledge. There is wide consensus that the ultimate, most highly prized goal of second language acquisition is spontaneous and unreflected L2 use, which is underlined by implicit knowledge (Sharwood Smith, 1981). However, the distinction between implicit and explicit knowledge has not been explicitly made in previous research that examines the influence of age and language aptitude. In addition to the consideration of implicit and explicit knowledge, the present study makes references to human learning mechanisms, that is, implicit and explicit learning. Individual difference variables, namely age and language aptitude, are selected as independent variables, and their predictive power for L2 outcomes is examined.

The examination of implicit and explicit knowledge provides an opportunity to reconsider the influence of age and aptitude. Following DeKeyser (2000) and DeKeyser et al.'s (2010) hypothesis, age is considered to be only relevant to the implicit learning system. This is in line with the maturational constraints view of age

effects, which acknowledges the influence of the biological factor age but rejects the traditional view of critical/sensitive period(s). Language aptitude, on the other hand, has been suggested as a mediating factor of the negative influence of age, and it has been hypothesised that language aptitude mainly functions in the explicit learning system. The extant research findings regarding age-aptitude-L2 outcomes are mixed, and the present study is conducted to examine this issue further.

As a study that investigates the age-aptitude interaction in relation to implicit and explicit L2 knowledge, it brings implications for the following aspects. Firstly, distinguishing between implicit and explicit knowledge is a relatively new approach to examining language proficiency, and this would add to current research in the area of implicit and explicit knowledge. Secondly, the study of implicit and explicit knowledge will provide information about different learning mechanisms. Although it is not possible for the present research to address the interface issue of implicit and explicit knowledge directly (see Section 2.2.3), some insights into how age and aptitude work for implicit and explicit learning can be drawn. Last but not least, the study of age and aptitude interaction will add to the current research in age effects and language aptitude. It is believed that the present study will give information about the form of age effects, that is, whether the age influence is manifested as a finite period or a linear decline, as well as information about the way aptitude functions in L2 development.

1.6 Thesis outline

The research context, personal incentives, basic terminology distinctions, and rationale have been presented so far. Following this chapter, a review of literature will be presented in Chapter 2. Relevant literature on implicit and explicit learning, and implicit and explicit knowledge will be reviewed first. Then, literature on the influence of age in second language acquisition will be reviewed. Influential theories, as well as empirical evidence, are discussed in this part. Then, literature on the role of language aptitude in second language acquisition is reviewed, including a

discussion of the construct and components of language aptitude, and relevant empirical research. At last, studies that addressed the age-aptitude relationship with L2 outcomes are reviewed in detail. Respective research questions are also presented in this chapter.

Chapter 3 presents the methodology, which includes a description of the research participants, the target structures, and the instruments used. The reliability of the measures in the present study is also examined in this chapter. The pilot study is presented at the end of this chapter.

In Chapter 4, the results of the present study are presented. This chapter starts with descriptive statistics results and some detailed discussion on aspects of the linguistic measures used. Construct validation results of the L2 outcomes measures and language aptitude measures are also presented. Then, results of the inferential statistics conducted in answering the four research questions are presented. Lastly, data collected from the questionnaire is shown at the end of this chapter.

Results presented in Chapter 4 are discussed in Chapter 5. This chapter starts with a discussion of the validity of the measures used, and then proceeds to discuss the extent to which hypotheses made with regard to research questions are confirmed or rejected.

The thesis concludes with Chapter 6, in which the main findings are summarised. Implications, limitations, and further research directions are discussed in this chapter.

CHAPTER TWO

LITERATURE REVIEW

In this chapter, relevant literature is reviewed. The chapter starts with a discussion of the implicit/explicit distinction in second language acquisition, followed by a review of the extant literature on age and language aptitude. The relevance of age and language aptitude to implicit and explicit knowledge is also addressed.

2.1 Cognitive accounts of second language acquisition

By taking a cognitive stance towards second language acquisition, the acquisition processes and the representation of L2 knowledge are attributed to the involvement of mental processes (R. Ellis, 2008b). The development of this cognitive view originates mainly from cognitive psychology, in which both the notion of implicit and explicit L2 knowledge and language aptitude have been investigated. The central claim of a cognitive view is that language acquisition is similar in nature to any other kind of learning in drawing on a common set of cognitive processes, namely implicit and explicit learning.

According to a cognitive perspective, L2 knowledge is represented as implicit and explicit knowledge, and the corresponding cognitive processes are implicit and explicit learning respectively. However, it should be noted that this kind of distinction is one way of seeing learning from a cognitive perspective, and other ways of interpreting L2 knowledge and L2 processing also exist. The implicit versus explicit distinction is discussed here because it lays the foundation of this study.

2.2 The implicit and explicit distinction in second language acquisition

2.2.1 Implicit and explicit learning

Implicit/explicit learning and implicit/explicit knowledge need to be differentiated. To put it simply, the former refer to the processes of learning, while the latter represent the end products of learning (R. Ellis et al., 2009; Schmidt, 1994). The following sections will start off by considering the important role of consciousness, and then definitions of implicit and explicit learning will be addressed. Different views of learning mechanisms will also be reviewed.

To define implicit and explicit learning, it is necessary to look at the construct of consciousness which is of central importance. In Schmidt's (1990; 1994; 2001) view, consciousness can be interpreted as intentionality, attention, control and awareness. Consciousness as intentionality is concerned with whether learning happens with or without deliberate planning. Conscious learning can happen during intentional (i.e. explicit) learning as well as during incidental learning. Under incidental learning conditions, the main objective is to learn one aspect of language, for example grammar, but the result might be the acquisition of another aspect, for example, new vocabulary. Nonetheless, consciousness is involved in this process. Consciousness as attention and control can be viewed in accordance to their relation to input processing and output processing. Attention in input processing refers to the noticing of information, while control in output processing refers to the retrieval of information. The control aspect of consciousness is closely related to automaticity, that is, spontaneous and fluent language production which is unconscious and does not involve conscious retrieval of explicit knowledge at all. Consciousness as awareness includes awareness as noticing and metalinguistic awareness. Awareness as noticing is mainly evident in the aspect of perception, and a certain degree of this kind of awareness is always involved in language learning. In contrast, metalinguistic awareness goes beyond perception, and involves analysis and reference to explicit grammatical rules.

On the basis of the consideration of consciousness, explicit learning is defined as a process that is generally both conscious and intentional (R. Ellis et al., 2009), it is a "conscious, deliberative process of concept formation and concept linking" (Hulstijn, 2002, p. 206). However, the definition of implicit learning is to some extent controversial. Some researchers argue that complete implicit learning does not exist since noticing with some degree of awareness cannot be avoided (e.g. Schmidt, 1994), while others maintain that complete implicit learning (i.e. noticing without awareness) is possible (N. C. Ellis, 2005; Williams, 2005). On the other hand, opinions about metalinguistic awareness are more unanimous, and all theorists would agree that metalinguistic awareness is not involved in implicit learning. Therefore, implicit learning is better defined as learning without metalinguistic awareness, and it refers to "the learning which takes place incidentally, in the absence of deliberate hypothesis-testing strategies, and which yields a knowledge base that is inaccessible to consciousness" (Shanks, 2003, p. 11).

The study of implicit and explicit learning has important implications for explaining the acquisition mechanisms of second language acquisition. However, both in the field of psychology and language acquisition, scholars disagree with regard to the roles of different learning mechanisms. Cognitive psychology researchers have different views regarding the existence of the implicit learning system and the way that this system is explained. Some researchers, such as Reber (1976), Wallach and Lebiere (2003), argue for a dual learning system, which states that human cognition is made up of an implicit system and an explicit system. The postulation of this kind of argument is on the basis of the differentiation between procedural memory and declarative memory, in which two distinct kinds of knowledge, namely productions (rules) and chunks (factual knowledge) are stored respectively. In contrast, Shanks (2003) contends that researchers are misguided when looking for any dissociation between implicit and explicit learning since both functional and neuro evidence is not convincing enough to establish such a position. Alternatively, a single learning system has been proposed and the observed

difference in performance is attributed to the difference in the retrieval processes of knowledge.

For second language acquisition researchers, the existence of a dual learning system is widely accepted. Dating back to the 1980s, Krashen's (1981) monitor theory distinguishes between subconscious language acquisition and conscious language learning, which can be seen as a recognition of the existence of two different acquisition mechanisms. Krashen also points out that the acquisition-learning distinction is not new; earlier researchers such as Corder (1967), Lawler and Selinker (1971), have suggested that two distinct types of cognitive processes work in synergy in the development of L2 proficiency. The use of consciousness in Krashen's theory has been criticised as too vague a term and the acquisition-learning distinction has been criticised as not being falsifiable. However, after Schmidt's (1990; 1994; 2001) detailed definition and discussion of the role of consciousness, it is considered that Krashen's overall proposal of a dual mechanism comprised of an implicit and an explicit learning system is likely to have some validity.

However, with developments in neuro-psychology, especially after the increasing use of hemodynamic technology (e.g. ERP, PET and fMRI), mixed findings regarding the relationship between implicit and explicit learning have been presented and a number of theories have been proposed. Some theories tend to support the dual learning system hypothesis, including the Declarative/Procedural model (DP model) (Ullman, 2001a; 2001b; 2005) and Paradis' model (1994; 2004; 2009); while others provide evidence favouring the single system hypothesis, such as the Competition model (Hernandez, Li, & MacWhinney, 2005; Hernandez & Li, 2007; MacWhinney, 2005; 2007), and the convergence hypothesis (Abutalebi & Green, 2007; Abutalebi, 2008; D. Green, 2003).

The DP model and Paradis' model share some similarities with the above-mentioned cognitive psychological view of implicit and explicit learning, for example, the belief in a dissociation of the procedural and declarative memory

systems. That is, the declarative memory system is mainly responsible for conscious explicit learning, and the procedural memory system underlines implicit learning for sequence and rules. Both of these models consider that the implicit system is available but somewhat attenuated as learner grows older, and also that declarative knowledge can be proceduralised to become implicit knowledge. Proceduralisation in the DP model is influenced by multiple intrinsic and extrinsic factors, such as L2 practice, the type of input received; while for Paradis, the key is to "have repeated practice (involving both comprehension and production) in interactive communicative situations" (Paradis, 2009, p. 4). However, in terms of the roles of the explicit system, the two models give different predictions: the declarative memory system is implicated both in implicit and explicit learning in the DP model, while it is claimed to be only relevant to explicit learning in Paradis' model.

In contrast to both the DP and Paradis' models, the competition model and the convergence hypothesis argue for a unified system for first and second language acquisition. These two theories are more in line with Shank's (2003) view that there is a single knowledge source underlying acquisition and performance. The competition model is developed from the notion of Construction Grammar, which regards input as several subcomponents including cues, arenas, chunking, storage, codes, and resonance (Hernandez et al., 2005; MacWhinney, 2005; 2007). According to this model, there is no qualitative difference between first and second language acquisition in the sense that language acquisition involves the formation of associative maps in the brain to store linguistic forms, including syllables, lexical items, and constructions. The convergence hypothesis holds the same view regarding how L2 is acquired, but it claims that the representation of L2 gradually converges with that of L1.

It can be summarised that implicit learning is understood as an unconscious process, and explicit learning as a conscious process. Whether complete implicit learning (learning without consciousness at all) is possible is a question under debate, but current evidence suggests that implicit and explicit learning systems can be

viewed dichotomously. R. Ellis (2009) suggests that although some psychological and neurological evidence seems to imply the function of a single learning mechanism (i.e. no distinction between implicit and explicit learning in language acquisition), making a distinction between the two learning systems would be an appropriate starting point to gain deeper understanding of these learning mechanisms. Therefore, in line with the views of the majority of second language researchers, it is concluded that both implicit and explicit learning mechanisms are functional for second language acquisition.

2.2.2 Implicit and explicit knowledge: dichotomous or continuous?

The distinction between implicit and explicit knowledge provides a way to explain the representation of L2 knowledge. Following the dual learning mechanism hypothesis, implicit and explicit knowledge is assumed to be the product of implicit and explicit learning respectively. As there is considerable debate regarding implicit and explicit learning, it is probably not surprising to see some controversy regarding the degree of differentiation between implicit and explicit knowledge. Scholars who hold the dual learning mechanism view either consider implicit and explicit knowledge to be completely separate (Krashen, 1981), or at least agree that they are distinct (Paradis, 1994). The logic behind this view has again originated from evidence of implicit and explicit memory, which is postulated to be related to distinct neuroanatomical regions (Paradis, 1994; 2004). Implicit memory is composed of "a network of specific frontal, basal-ganglia, parietal and cerebella structures", whereas explicit memory depends on "the hippocampal region, entorhinal cortex, and peripheral cortex" (Ullman, 2004, pp. 235, 238). Further supporting evidence is drawn from studies of neurological impairment, which show that the damage of one memory system does not influence performances related to the other system (Ullman, 2001a; 2001b; 2005). For example, patients who suffer from Parkinson's Disease do not have problems in lexical processing (which rely on explicit memory), even

though they do have difficulties in grammatical processing caused by the damage in implicit memory (Schacter, 1992; Schacter, Chiu, & Ochsner, 1993; Tolentino & Tokowicz, 2011). Aphasia patients who have learnt their L2 formally in classroom settings may still be able to use this language which is assumed to be linked to their explicit memory when they have lost the ability to speak in their L1 (Paradis, 2004). In a similar fashion, Ullman's DP model argues forcibly about the distinctiveness of implicit and explicit memory, in which mental grammar and mental lexicon are stored respectively (2001a; 2001b; 2004).

In contrast, some other researchers view implicit and explicit knowledge as a continuum. In Dienes and Perner's (1999) view, explicit knowledge is "the representations of one's own attitude of knowing the fact", and implicit knowledge refers to "the representations that merely reflect the property of objects or events without predicating them of any particular entity" (p. 752). In this hypothesis, knowledge is taken as an attitude held towards a proposition, in which implicit and explicit knowledge has a partial hierarchical relation (i.e. if a higher aspect is known explicitly then each lower one must also be known explicitly), and the implicit/explicit distinction can be made at each level. However, this hypothesis has received many objections, and these objections state that it fails to demonstrate the continuum relation between implicit and explicit linguistic memory itself (R. Ellis, 2004), and to account for differences in learning performance under implicit and explicit conditions (Noelle, 1999). Those who hold this perspective are also criticised for the way that they define implicit and explicit knowledge, as R. Ellis (2004) noted, "explicit knowledge can be viewed as the outcome of an attitude but is distinct from it" (p. 229). In short, although doubts still remain, both in terms of the existence of implicit and explicit knowledge, and the relationship between the two, it is more reasonable to assume that a distinction can be made between implicit/explicit learning of an L2 and between implicit/explicit knowledge (R. Ellis, 2005; R. Ellis et al., 2009; Hulstijn, 2002; Krashen, 1981; Paradis, 1994). In other words, the dual learning mechanism hypothesis and the dichotomous view of implicit/explicit

knowledge are supported by current research findings.

2.2.3 The interface: how do implicit and explicit knowledge develop?

The interface issue has been a topic of controversy in the study of implicit and explicit knowledge for many years. This is concerned with the extent to which implicit knowledge converts into explicit knowledge and vice versa. An answer to this question would give indications of the relationship between implicit and explicit learning, and also of the role of explicit knowledge in the development of implicit knowledge. Three positions have been suggested, namely the non-interface position, the strong interface position and the weak interface position. These will be discussed in greater detail below.

The non-interface position holds that implicit and explicit knowledge is distinct, and there is no possibility of one type of knowledge transforming to the other. The implicit and the explicit systems are independent from each other, and the development of each is correspondingly independent as well. Rules can be acquired implicitly, or learnt explicitly, and this is reflected in learners' performance as acquired competence and learnt competence, respectively (Krashen, 1981). This position is based on the dual learning mechanism hypothesis which involves two distinct learning mechanisms (Hulstijn, 2002; Krashen, 1981), and correspondingly two memory systems which store the acquired knowledge (Paradis, 1994). The different means of retrieving knowledge, namely by controlled or automatic processing (R. Ellis, 1993; 2004; 2009), have also been taken as supporting evidence for this position. In other words, when practice is defined as the specific activities being engaged in as a means to develop knowledge and skills of the L2 (DeKeyser, 2007), the non-interface position holds that practice has no role to play in helping explicit knowledge convert to implicit knowledge.

In contrast, the strong interface position states that implicit and explicit knowledge can convert into each other. Implicit knowledge can be transferred to explicit knowledge by means of conscious reflection and analysis, and explicit

knowledge can become implicit knowledge through practice (Sharwood Smith, 1981; DeKeyser, 1998; 2007). Central to this position is the recognition of the role of practice in converting explicit knowledge to implicit knowledge. Accordingly, language instruction is viewed as playing a two-fold role in language acquisition. First, instruction provides learners with conscious rules, that is, explicit knowledge. Second, instruction provides learners with opportunities to apply the rules through practice, and it is believed that explicit knowledge will gradually turn into implicit knowledge. In other words, the strong interface position posits that explicit knowledge can be proceduralised into implicit knowledge given a sufficient amount of practice.

Another different account, somewhere in between the non-interface and the strong interface positions, is the weak interface position (N. C. Ellis, 1994; R. Ellis, 1993; 1994). This position acknowledges the possibility of explicit knowledge converting to implicit knowledge, but also suggests certain conditions as prerequisites. There are three different versions of this position, depending on the different theories being drawn on.

The first version is based on the learnability/teachability hypothesis proposed by Pienemann (1989), which emphasizes the readiness of learners. That is, teaching is only effective when learners are psycholinguistically ready to acquire the particular linguistic structure. On the one hand, this position divides learning into different stages, and underlines the order of acquisition. For example, a structure of Stage X+3 cannot be acquired by learners at Stage X, because they are not yet ready for this structure. Only learners at Stage X+2 are ready for structures of Stage X+3. The interpretation that follows is that "teaching is impossible since L2 acquisition can only be promoted when the learner is ready to acquire the given items in the natural context" (Pienemann, 1989, p. 61). However, on the other hand, this position also emphasizes that teaching can facilitate learning for learners who are ready for it (Pienemann, 1984). Therefore, this version of the weak interface position postulates a developmental constraint for explicit knowledge to be converted to implicit

knowledge. However, it should be noted that this hypothesis does not address the difference between implicit and explicit knowledge directly, and the focus of this theory mainly lies on the developmental sequence of language acquisition.

The second and third versions of the weak interface position are partly similar to the claims from the strong interface position in the sense that explicit knowledge is considered to be contributing to the acquisition of implicit knowledge. Yet, this contribution is indirect as opposed to the direct contribution proposed by the strong interface position. N. Ellis's (1994) version of the weak interface position suggests that implicit knowledge can only be developed by implicit learning, but explicit knowledge (and also explicit learning) can support the implicit learning process. This facilitative role of explicit knowledge is achieved by making relevant pattern or chunk features salient for learners to be consciously aware of in their working memory, so that instruction consisting of declarative rules can facilitate the acquisition of implicit knowledge. In other words, N. Ellis (1994) contends that both implicit and explicit learning processes are dynamically involved in second language acquisition, but explicit knowledge only plays an indirect role for the development of implicit knowledge. In his view, explicit knowledge cannot be converted into implicit knowledge.

The third version, proposed by R. Ellis (1993; 1994), similarly promotes the view that explicit knowledge can facilitate the development of implicit knowledge. Explicit knowledge facilitates the noticing of linguistic structures and the noticing of the gap between input and output, which in turn promotes the development of implicit knowledge. However, not all L2 knowledge that originates in an explicit form can be converted into implicit knowledge. For example, certain grammatical structures such as negation and third person singular can only be converted into implicit knowledge when the learner has reached a stage of development that allows for the integration of the structure into the interlanguage system. In other words, the conversion of explicit knowledge into implicit knowledge is circumscribed by developmental constraints because grammatical features can only be acquired in a

fixed sequence. However, later in his work, R. Ellis (2006; 2008a) tends to place more emphasis on the indirect contribution of explicit knowledge by means of helping the learners to notice linguistic forms and then to make comparisons between noticed forms and their own interlanguage.

The present study is not able to directly address the interface issue because only the final product of second language acquisition is examined. However, since distinctions are made between implicit and explicit learning, and between implicit and explicit knowledge, it is assumed that implicit learning would primarily lead to implicit knowledge, while explicit learning would lead to explicit knowledge. In saying so, the present study does not exclude the possibility of explicit knowledge contributing to implicit knowledge.

2.2.4 Definition and measurement of implicit and explicit knowledge

One's repertoire of linguistic knowledge is comprised of a combination of implicit and explicit knowledge, but it is implicit knowledge that underlines linguistic competence. Both scholars who argue for the important role of Universal Grammar and those who argue for the utilization of a simple cognitive mechanism (Gregg, 2003) agree on this position. Learners can potentially develop implicit and explicit knowledge of any aspect of a language, including grammar, pronunciation, vocabulary and others. However, implicit and explicit grammar knowledge is the most widely studied linguistic aspect to date. There are multiple reasons for this phenomenon. Firstly, grammar plays a central role in the acquisition process of L2 learners, especially those who learn the language in a formal instructed context. Secondly, grammar, which consists of linguistic rules, is more accessible to conscious reflection compared to vocabulary or pronunciation (Odlin, 1989); so the testing of grammar is more practical and feasible than the testing of other linguistic aspects. Thirdly, considering the long history of linguistic research, many pre-existing methods of assessing grammar have been established and validated.

Therefore, it is not surprising that the discussion of implicit and explicit knowledge is mainly related to grammar.

In a series of publications, R. Ellis (1994; 2004; 2005; 2009) provides a detailed description of the characteristics that are hypothesised to define and differentiate implicit and explicit knowledge. Table 2.1 is reproduced from R. Ellis (2008, p. 418) and summarises these characteristics.

Table 2.1 Characteristics of implicit and explicit knowledge

Characteristics	Implicit knowledge	Explicit knowledge
Awareness	Learner is intuitively aware of linguistic norms.	Learner is consciously aware of linguistic norms.
Type of knowledge	Learner has procedural knowledge of rules and fragments.	Learner has declarative knowledge of rules and fragments.
Systematicity	Knowledge is variable but systematic	Knowledge is often anomalous and inconsistent.
Accessibility	Knowledge is accessible by means of automatic processing.	Knowledge is accessible only through controlled processing.
Use of L2 knowledge	Knowledge is typically accessed when learner is performing fluently.	Knowledge is typically accessed when learner experiences a planning difficulty.
Self-report	Non-verbalizable.	Verbalizable.
Learnability	Potentially only learnable within the critical period.	Learnable at any age.

Researchers have endeavoured to empirically operationalise some of the above characteristics, and accordingly tests have been devised to independently access implicit and explicit knowledge (R. Ellis, 2004; 2005; R. Ellis et al., 2009). In the following paragraphs, the measurement of implicit and explicit knowledge is reviewed in detail.

In terms of what is known about how implicit and explicit knowledge can be accessed, tests that involve automatic processing are considered as likely measures of implicit knowledge whereas tests that involve controlled processing measures of

explicit knowledge. Automatic processing means that unplanned language use has taken place, so that implicit knowledge that is tacit and procedural in nature is drawn on. In comparison, controlled processing requires the involvement of attentional processes, so that explicit knowledge is drawn on (R. Ellis et al., 2009). However, it has been suggested that it might not be possible to empirically distinguish implicit and explicit knowledge given that explicit knowledge may be accessible in unplanned language use as well (DeKeyser, 2003; Williams, 2009). DeKeyser (2003) argues that there is a possibility that explicit knowledge can be proceduralised to the extent that it is functionally equivalent to implicit knowledge, so that both implicit and explicit knowledge may be available for automatic processing (N. C. Ellis, 1994; Rebuschat, 2013). However, the extant empirical findings appear to suggest that implicit and explicit knowledge are distinguishable, as long as tests that target different types of knowledge are carefully designed and administered (Bowles, 2011; R. Ellis, 2005; R. Ellis et al., 2009; Han & Ellis, 1998; Zhang, 2015).

To create testing conditions that require automatic or controlled processing, researchers have 1) manipulated the time available to participants, and 2) utilised different testing modalities (i.e. stimuli presented aurally or visually). It is hypothesised that tests that impose a time pressure would tap implicit knowledge, whereas tests that do not impose a time pressure would allow for the use of explicit knowledge. This is because participants are forced to engage in automatic processing and to utilise their available implicit knowledge in timed tests, while they could engage in controlled processing and resort to their explicit knowledge in untimed tests. In other words, the manipulation of time available to participants creates conditions for distinctive types of processing (automatic vs. controlled), thus making timed and untimed tests likely measures of implicit and explicit knowledge respectively. Several empirical studies have provided evidence in support of this operationalisation. For example, Ellis and Loewen (2007) developed a timed and an untimed grammaticality judgement test (TGJT and UGJT, respectively), and their factor analysis results demonstrated that these two tests measured distinctive

constructs that they labelled as implicit and explicit knowledge (Loewen, 2009). The manipulation of time for the purpose of measuring different types of knowledge also received support from subsequent studies (Bowles, 2011; Godfroid, et al., 2015; Kim & Nam, 2017; Zhang, 2015).

In addition to imposing a time pressure in the grammaticality judgement tests, researchers have argued that tests that are presented aurally are more likely to tap implicit knowledge than tests that are presented visually (Kim & Nam, 2017; Spada et al, 2015). This is probably because the presentation of aural stimuli places demands on one's short-term memory and working memory, which are limited in capacity (Wong, 2001). When test stimuli are presented aurally, participants would need to use their implicit knowledge because "the signal is fleeting and not available for prolonged reflection" (Rebuschat, 2013, p. 612), so that it is difficult to pay attention to both form and meaning in the processing of aural input. In contrast, when written stimuli are presented, the visual information that participants see would probably help them with the processing (and also reprocessing) of the stimuli. Therefore, tests that are timed and aurally presented have a higher probability to access implicit knowledge.

Another aspect that could differentiate the use of implicit and explicit knowledge in grammaticality judgement tests is related to the level of awareness, that is, whether participants are intuitively or consciously aware of linguistic norms during test completion. Knowing intuitively that a sentence is correct or incorrect implies that implicit knowledge is probably being used in making judgements. In contrast, drawing consciously on one's knowledge to decide whether a sentence is correct or incorrect implies that explicit knowledge is most likely being tapped. This conscious attention to one's knowledge suggests that participants are using their explicit knowledge of grammar to analyse the stimuli so as to decide their grammaticality. To operationalise this characteristic, Loewen (2009) asked a question to establish the level of awareness for each sentence in the untimed grammaticality judgement test. Participants were required to indicate on what basis

they made their grammaticality judgement, whether by "rule" or by "feel". The selection of "rule" in untimed grammaticality judgement tests is believed to indicate that explicit knowledge is used, while the selection of "feel" indicates that implicit knowledge is probably involved.

It is believed that there are differences in terms of whether each type of knowledge can be reported or verbalized by learners. While implicit knowledge is only evident in learners' verbal behaviour, explicit knowledge is potentially verbalizable. In other words, having implicit knowledge of a certain structure means that one would be able to use the structure correctly in communication, although he or she may not know the underlying rules of the usage. In comparison, having explicit knowledge would mean that some explanation could be provided regarding the rules of the usage, and this is often referred to as the verbalization of explicit knowledge. Verbalization can take the form of non-technical explanations, or technical terms, that is, metalanguage. For example, in identifying the problem in the following sentence "He play soccer very well.", a non-technical explanation would be "an *s* must be put after play because the sentence is talking about he", while an explanation using metalanguage would be "play need to be plays to indicate the third person singular". Although metalanguage need not be used to verbalize explicit knowledge, it has been suggested that metalanguage is closely related to explicit knowledge (Elder, 2009; R. Ellis, 2009).

Elder (2009) developed a metalinguistic knowledge test that requires participants to use their knowledge of English metalanguage (see Section 3.4.3.2 for a detailed description of this test). Undoubtedly, tests that tap metalanguage would reflect participants' level of explicit knowledge, but the testing of explicit knowledge does not necessarily require the verbalization of explicit knowledge (i.e. to provide explanations of grammar rules). As discussed above, the untimed grammaticality judgement test developed by Loewen (2009) has been suggested as a valid measure of explicit knowledge and this test does not require rule explanation at all. Rather, a testing condition that predisposes participants' attention to linguistic forms was

created, so that participants would be encouraged to make conscious reflection on their explicit knowledge. In comparison, the involvement of metalanguage in Elder's (2009) metalinguistic knowledge test would require a higher degree of additional attention to, and conscious reflection of grammar rules. In this sense, metalinguistic knowledge tests are probably purer measures of explicit knowledge than untimed grammaticality judgement tests.

Implicit and explicit knowledge also differ with respect to learnability. It has been argued that implicit knowledge could only be acquired within the critical period, whereas explicit knowledge can be learnt at any age. The learnability difference is a reflection of the hypothesised difference between implicit and explicit learning mechanisms, and it is also the characteristic that connects the study of implicit and explicit knowledge with the study of individual differences. Implicit knowledge is learnable, but age seems to be the single most influential factor that restricts learners' ability to learn implicitly (Birdsong, 2006). The influence of age is evident given that very few L2 learners achieve native-like proficiency in naturalistic learning contexts. In contrast, learners who learn the L2 in instructed contexts normally have a large amount of explicit knowledge, yet they might still encounter difficulties in spontaneous language production.

This difference in learnability indicates that one way of examining the validity of measures of implicit and explicit knowledge is to compare the performances of second language learners with those of native speakers. This was the approach adopted by R. Ellis and his colleagues (R. Ellis et al., 2009), as well as by some other researchers (Bowles, 2011; Zhang, 2015). Research findings have repeatedly shown significant differences between the implicit and explicit knowledge of L2 learners and native speakers. L2 learners generally show a lack of implicit knowledge in comparison with native speakers, but they tend to have more explicit knowledge than their counterparts. This is explicable, because the level of implicit knowledge could be very much influenced by the maturational factor of age, while explicit knowledge tends to be less influenced by this factor (see Section 2.3 for

more detail). Due to variations in learning age, it is expected that there are considerable variations in the levels of implicit knowledge developed by L2 learners. In comparison, the levels of implicit knowledge of native speakers are more homogeneous, as people generally have mastery of their first language. Thus, when comparing the levels of implicit knowledge between the two groups, L2 learners and native speakers have been identified as two distinctive populations.

Following attempts to define and characterise implicit and explicit knowledge, a battery of tests was devised and validated by R. Ellis and his colleagues. They developed an oral production test, an oral elicited imitation test, and a timed grammaticality judgement test as measures of implicit knowledge; an untimed grammaticality judgement test and a metalinguistic knowledge test as measures of explicit knowledge (R. Ellis et al., 2009; Erlam & Akakura, 2016). Similar tests have been used by later researchers (Bowles, 2011; Kim & Nam, 2017; Zhang, 2015), and they have concluded, in line with R. Ellis et al. (2009), that this battery of tests shows some evidence of validity.

However, a number of recent studies have challenged the validity of some of the above tests. For example, Suzuki and DeKeyser (2015) argue that elicited imitation tests are not valid measures of implicit knowledge, because they found a correlation between the scores of an elicited imitation and a metalinguistic knowledge test. Suzuki (2017), and Vafaee, Suzuki and Kachisnke (2017) present other evidence that casts some doubt on the validity of the test battery developed by R. Ellis et al. (2009). They argue that R. Ellis et al.'s (2009) tests are not sensitive enough to distinguish between implicit knowledge and automatized explicit knowledge. In their view, measures of reaction time, such as self-paced reading and word monitoring tests, are better measures of implicit knowledge (see Section 5.1 for detailed discussion).

2.2.5 Research question one

The present study follows R. Ellis et al. (2009) and makes modifications to their

battery of tests. The adapted measures are explained in more detail in Chapter 3. Also in line with previous research, the present study recruits two groups of participants, one being L2 learners (i.e. the non-native speakers; NNS) and the other speakers for whom English is a L1 (i.e. the NS). Here, in the light of previous research findings, the first research question is asked. This question asks about the profiling of the implicit and explicit knowledge of the NNS and the NS, and if there is a difference between the implicit and explicit knowledge of the NNS in comparison to the NS. Four hypotheses are made:

Hypothesis 1a: The NNS will have more explicit knowledge than implicit knowledge of English.

Hypothesis 1b: The NS will have more implicit knowledge than explicit knowledge of English.

Hypothesis 1c: The level of implicit knowledge of the NNS will be lower than that of the NS.

Hypothesis 1d: The level of explicit knowledge of the NNS will be higher than that of the NS.

2.3 Age and second language acquisition

2.3.1 The profile of age effects: critical period or general maturational constraints?

The "younger is better" view in second language acquisition might be the perspective that is held by many laymen owing to both their personal observation of child learners with high achievements in L2 and the promotion of language courses in primary schools. A wealth of previous research has shown that the age factor figures prominently in second language acquisition with regard to ultimate language attainment. However, the question at issue to date is the nature of this influence of age.

One possibility is that the effects of age in second language acquisition are

manifested as a *critical period* (Lenneberg, 1967; Penfield & Robert, 1959) or several critical periods for different linguistic domains (DeKeyser, 2000; DeKeyser et al., 2010; Granena & Long, 2013a; Hyltenstam & Abrahamsson, 2003; Long, 1990). In Lenneberg's (1967) view, the critical period includes both an onset (age two) and an offset (puberty) of language acquisition. The onset age of two was believed to be the time when language development starts to take place, and the offset age of puberty was in line with the hypothesised completion of language lateralization. It should be noted that the critical period here is only presumed to be related to the aspect of pronunciation, which is believed to be directly initiated by neurophysiological mechanisms (Scovel, 1969). The lexical and syntactic patterns of language lack such a neuromuscular basis, and therefore only pronunciation is believed to be influenced by age (Lenneberg, 1967; Scovel, 1969; 1988; 2000). However, the notions "age of onset" and "age of offset" were rejected by later findings, which show that language development is a continuous process right after birth (Singleton & Ryan, 2004), and that language lateralization is completed by the age of five (Molfese & Molfese, 1997). Therefore, language lateralization is not a valid explanation for declining learning ability, and in turn the original proposal of the offset of puberty was questioned by researchers.

In addition to the rejection of a biological explanation of the *critical period hypothesis* (CPH), the terminology "critical period" also received some challenges. A number of scholars advocated using the terms *sensitive period* and *sensitive period hypothesis* (SPH) instead (Hess, 1975; Oyama, 1976; 1978), because they think that the influence of age in language acquisition happens gradually, and that the connotation of critical period is too absolute. The formulation of the notion of critical period seems to promote an all-or-nothing view of language acquisition after puberty, which is obviously not supported by empirical evidence. They contend that while the distinguishable onset and offset of the CPH represent a window of opportunity in second language acquisition, the SPH is less restrictive in terms of human learning capacity and predicts "a gradually declining effectiveness of the

peripheral input" after the offset of the sensitive period (Eubank & Gregg, 1999, p. 68).

Following the original formulation of the CPH (or the SPH), the spectrum of critical period or sensitive period research expanded to a much wider extent in the following decades. Findings have lent support to the initial separation of pronunciation from other linguistic domains, showing that pronunciation is indeed most affected by age. In addition, research has shown that the effects of a critical/sensitive period are not restricted to pronunciation; other linguistic domains such as morphology and syntax are also subject to the influence of age. However, it has been suggested that the influence of age on different linguistic domains is different, and that there might be multiple critical periods or sensitive periods (Long, 1990; Granena & Long, 2013a; Meisel, 2011; Seliger, 1978). Singleton and Ryan (2004) discussed the onset of different linguistic domains including phonology, grammar and lexicon in detail, and proposed multiple sensitive periods. The extant research findings suggest that L2 phonology, lexis and collocation are most influenced by age, and that the age period of 0 to 6 years is believed to be the critical/sensitive period for nativelike attainment in these domains (Hyltenstam, 1992). In comparison, L2 morphology and syntax are less influenced by age, and the critical/sensitive period is hypothesised to be much later, up until mid-teens (DeKeyser, 2000; Johnson & Newport, 1989; Patkowski, 1990).

Terminological discrepancies aside, the theoretical underpinning of the proposed critical/sensitive period or multiple critical/sensitive periods is similar, because a bounded period of time is believed to be the optimal time for second language acquisition. This postulated time period (or multiple periods) has the following characteristics: 1) an onset, 2) an offset (or terminus), and 3) a capacity to indicate ultimate L2 proficiency (Harley & Wang, 1997). The first two characteristics specify an observable advantageous period for acquisition, which implies a discontinuity in language acquisition before and after the offset. In other words, the onset age and the offset age mark the boundary of the critical/sensitive

period(s), with the former indicating the start and the latter indicating the end of this period(s). The offset or terminus feature also implies a point of change, which means a qualitative difference in ultimate attainment before and after the offset. To qualify as evidence in support of the critical/sensitive period hypothesis, empirical evidence needs to show different patterns in linguistic performance before and after the offset(s) of the critical/sensitive period(s).

Just as researchers disagree with regard to the offset of the critical/sensitive period(s), there is considerable disagreement in terms of the form of the critical/sensitive period(s). For example, Johnson and Newport (1989) believed that the age effect can be described as a *stretched Z*, which begins with a ceiling period (bounded period of time, and no observable age effects at the earliest ages), then followed by a decline (bounded, refers to the critical/sensitive period(s)), and at last flattened out (unbounded, the bottoming out of the effect of age). Another two possible forms of the effect of age are proposed as a *stretched L* and a *stretched 7*, as described in Birdsong (2006). The stretched L function starts with a bounded sloping period, followed by an unbounded flattening out period; whereas the stretched 7 starts with a bounded ceiling period, followed by an unbounded down slope. However, empirical studies fail to provide evidence for these hypothesised age functions. In a reanalysis of Johnson and Newport's (1989) data, the crucial feature of the flattening out in the stretched Z shape disappeared when the age cut-off point was moved to 20 years. Rather, the data showed a significant negative linear correlation between age and language performance (Bialystok & Hakuta, 1994).

Contrary to the critical/sensitive period(s), there is some empirical evidence that suggests a linear relationship between age and ultimate attainment, which shows an effect of age across the spectrum of all ages. The negative linear influence maintains as age increases, without an observable point of discontinuity (Bialystok & Hakuta, 1999; Flege, Yeni-Komshian, & Liu, 1999; Flege, 1999; Hakuta, Bialystok, & Wiley, 2003). This pattern of influence can be seen as some counter evidence for the hypothesised critical/sensitive period(s).

To reconcile these findings, some scholars suggest that it is more appropriate to accept the view that second language acquisition is maturationally constrained. The proposition of maturational constraints is a comprehensive notion that accounts for human language acquisition under all conditions, including first and second language acquisition (Long, 1990). This postulation covers all relevant research regarding age and human language learning ability, regardless of which specific theory is employed (Hyltenstam & Abrahamsson, 2003). Under this postulation, the potential causal relationship between the maturational factor of age and acquisition outcomes is well acknowledged, irrespective of whether the effect of age is manifested as a finite period(s) or a linear influence.

Research in first language acquisition also provides some evidence in support of the maturational constraints view of language acquisition. For instance, the existence of maturational constraints was confirmed by the study of delayed L1 acquisition (Curtiss, 1977; 1988). The example given is that of Genie, who was deprived of linguistic input before the age of 13 and was only able to reach the level of proficiency similar to a two-year-old. It is therefore argued that the human brain is adjusted for language acquisition at the early years of life, and that this will lead to inevitable success in L1 acquisition given normal exposure to the language. In contrast, the brain seems to lose the ability of language acquisition as a result of maturation in later years of life, which results in unsatisfactory outcomes, as is the case of Genie. Such findings from delayed L1 acquisition lends direct support to the hypothesis that humans are adapted for language acquisition at an early age, but as maturation takes place (i.e. age increases), one's ability to naturally acquire a language is somewhat affected.

Similarly, research in second language acquisition that shows an impact of age on ultimate attainment can be taken as evidence in support of the maturational view. The extant research findings have repeatedly shown negative correlations between age and ultimate attainment, which is adequate enough to demonstrate that the outcomes of second language acquisition are constrained by the maturational factor

of age (e.g. Abrahamsson & Hyltenstam, 2008; 2009; DeKeyser, 2000; DeKeyser et al., 2010; Granena & Long, 2013a; Hyltenstam & Abrahamsson, 2000; Hyltenstam & Abrahamsson, 2001; Johnson & Newport, 1989). In addition, the generally observed big variation in the L2 outcomes of adult learners provides some further evidence that the maturational factor age plays a role in second language acquisition. It is likely that age has a stronger influence on second language acquisition than it does on first language acquisition.

It can be summarised so far that age has a profound influence on second language acquisition. Although researchers differ in their views regarding whether the effect of age is manifested as (a) critical/sensitive period(s) or a linear influence, there is general consensus that second language acquisition is maturationally constrained.

2.3.2 Causes of maturational constraints

As reviewed above, the biological explanation of age effects as caused by brain lateralisation (Lenneberg, 1967) has been falsified. Researchers have suggested other explanations to account for age effects in second language acquisition.

Some scholars have examined the possibility that other psychological or social factors, such as motivation and input that co-vary with age could be the cause of age effects. For example, Bialystok and Hakuta (1999) hypothesised that cognitive factors and age are both possible candidates that influence ultimate attainment, and that the influence of cognitive factors could be considered as significant when the influence of age is partialled out. However, it is unlikely that differences in motivation can explain age effects, because such evidence has not been found in a number of studies (Johnson & Newport, 1989; Oyama, 1976; 1978), and it is probably incorrect to assume that children and young learners are more motivated than adults (Long, 1990). Likewise, input alone cannot offer satisfactory explanation for age effects, because if learners have continued access to L2 input of high quality

(i.e. from native speakers), it should be possible for adults to attain a native-like accent in spite of a late start (Klein, 1995). However, it is observed that adult learners who successfully reach a native-like level of attainment are exceptional cases.

It is apparent that maturational constraints on second language acquisition are too robust to be easily accounted for by individual psychological or social factors. It is more likely that variations in ultimate outcomes of second language acquisition are a result of the interplay between maturational and non-maturational factors, and the non-maturational factors probably create advantageous circumstances which facilitate L2 acquisition (Abrahamsson & Hyltenstam, 2008; Bongaerts, Planken, & Schils, 1995). One factor that has been given prominence is language aptitude, that is, the cognitive ability to acquire a second language. The relevance of this factor is discussed in Section 2.4.

Another alternative explanation of age effects originates from consideration of the human cognition system. The general idea is that children and adults rely on different mechanisms for language acquisition. Unlike children who learn implicitly, older learners and adults tend to rely on the more explicit and general problem-solving mechanism for language acquisition. The fundamental difference hypothesis (Bley-Vroman, 1988; 1989; 2009) presents a version of this hypothesis. It predicts that adults differ from children in that they no longer have access to the innate device of language acquisition (i.e. universal grammar), so they have to rely on the general problem-solving mechanism. A slightly different view is put forward by the competition hypothesis (Felix, 1985), which assumes that the implicit learning mechanism of adults competes with their explicit learning system, but that it loses power eventually. Literature on statistical learning provides another perspective in understanding the possible deployment of the implicit learning mechanism in language acquisition. Statistical learning is viewed as a type of learning that is domain-general, which occurs unconsciously and automatically when learners are exposed to linguistic input (Rebuschat & Williams, 2013). Findings from recent

psychological studies of adult L2 learners show that these learners can learn from implicit or incidental learning conditions, which indicates that the implicit learning mechanisms might still be available for use regardless of learners' age (Leung & Williams, 2011; Simon, Howard & Howard, 2011; Williams, 2005). Under incidental learning conditions, adult learners are found to be able to acquire form-meaning connections and L2 syntax without intending to do so.

It has been pointed out that consideration of different learning mechanisms in language acquisition does figure in the CPH (Lenneberg, 1967), but has been to some extent misunderstood and overlooked by scholars (DeKeyser & Larson-Hall., 2005). Lenneberg's (1967) original definition of the CPH refers to the "automatic acquisition from mere exposure that seems to disappear after this age (puberty)" (p.176). It can be seen from this statement that CPH is supposed to be related to implicit learning only, which applies to both L1 and L2 acquisition. However, empirical research that aimed at verifying the CPH or SPH usually ignored the implicit and explicit distinction in learning conditions (i.e. implicit or explicit learning), as well as the nature of the outcome of the acquisition (i.e. whether implicit or explicit knowledge is acquired). It would not be surprising that learners who learn primarily in an instructed context have high ability in explicit knowledge, such as rule explanation in grammar. In contrast, those who learn in a naturalistic context should be able to show some degree of implicit knowledge.

Indeed, if implicit and explicit learning mechanisms are considered, the issue of language exposure becomes very important because it is related to the type of learning that might be involved in second language acquisition. In relation to instructed and naturalistic L2 learning contexts, the role of language exposure has been discussed. For example, Jia and Fuse (2007) conclude that language environment is a stronger predictor of performance variance than age (this experiment did not control for length of residence, which means that experimental data obtained might speak to learning rate, rather than learners' ultimate attainment). Similarly, Johnson and Newport (1989) measured age of arrival in the L2 country as

the crucial age factor, and they did not account for age of first exposure to English in their home country. The reason is that "the learning which occurs in the formal language classroom may be unlike the learning which occurs during immersion, such that early instruction does not necessarily have the advantage for ultimate performance that is held by early immersion" (Johnson & Newport, 1989, p. 81). In other words, by considering different learning contexts and making assumptions about the kind of learning that occurs in different contexts, age is assumed to have different roles in the implicit and explicit learning systems. Recent empirical results corroborate this idea, showing that age only has an influence on the implicit learning system (Granena, 2013b; 2014; Granena & Long, 2013a).

By assuming maturational constraints only influence the implicit learning system, the following explanation of age effects can be made. On the one hand, the lower likelihood that adult learners succeed in second language acquisition compared to child learners is because their implicit learning system is not available for them to use, so they have to rely on the less effective explicit learning system. On the other hand, there could still be a possibility for adult learners to reach at least near-nativeness in their L2. This may be because other cognitive factors, such as language aptitude, play an ancillary role in the explicit learning system, therefore making a high level of L2 proficiency possible. Current research findings support such an explanation to some extent: the extremely successful L2 learners reported in several studies were found to have high language aptitude (e.g. Ioup, Boustagui, El Tigi, & Moselle, 1994; White & Genesee, 1996), but there was still a general negative correlation between age and L2 proficiency (Bialystok & Hakuta, 1994; 1999; Birdsong, 1999; Flege et al., 1999). Equally intriguing is the fact that there are also cases when some learners fail to reach nativelike proficiency even if they start very early (Bialystok & Miller, 1999; Butler, 2000; DeKeyser, 2000; Flege et al., 1999; Hyltenstam, 1992), which is probably an indication of the involvement of other variables in implicit learning.

Without reference to specific learning contexts, that is, whether second

language acquisition happens only in a naturalistic context, for example for immigrants who only start acquiring the L2 after immigration, or a combination of learning in both an instructed and a naturalistic context, such as those who have the experience of learning an L2 in an instructed context and then immigrate to the L2 country; the investigation of maturational constraint seems to be limited in scope. However, previous research generally examined the influence of age in naturalistic contexts and did not investigate the knowledge resulting from learning in instructed contexts. In other words, in order to understand how individual difference factors influence L2 acquisition outcomes, there is a need to consider the learning contexts of learners, as well as the type of learning processes that are involved, including both implicit and explicit learning.

Nonetheless, the extant research findings imply that variation in L2 outcomes is influenced by the variables of age, cognitive factors, and their relevance to implicit and explicit learning. For example, Granena (2012) examined L2 learners' linguistic knowledge with the use of a battery of tests, which could be seen as measures of implicit and explicit knowledge. Granena (2012) reports that age has an influence on implicit knowledge based on her findings that even the early L2 learners (i.e. age of immersion between 3 to 6 years) differed significantly from the native speakers. In the light of previous findings, it is therefore predicted that given the same naturalistic learning context, a negative correlation could be found between age and the amount of implicit knowledge obtained as age increases.

2.3.3 Instances of adult nativelike learners

Several scholars argue that studies such as Johnson and Newport (1989) and DeKeyser (2000) do not address the issue of whether late L2 learners can ever attain full proficiency in an L2, that is, a level of proficiency that is indistinguishable from native speakers (Abrahamsson & Hyltenstam, 2009; White & Genesee, 1996). Nativelike attainment, if possible, must be rare and commonly used

random-sampling strategies might not even include such nativelike participants. In other words, although there is some evidence that second language acquisition is constrained by age, the absolute potential of late learners needs to be investigated by specifically targeting L2 learners who are known to be highly proficient to see if they have attained nativelike proficiency. Following this logic, scholars have endeavoured to empirically test the proficiency of some potential nativelike participants and then to compare their proficiency to that of native speakers (e.g. Abrahamsson, 2012; Abrahamsson & Hyltenstam, 2008; 2009; Birdsong, 1992; Coppieters, 1987).

Studies that focus on potential nativelike speakers share the common feature of employing some sort of screening procedure for the selection of participants. For example, Coppieters (1987) selected 21 successful adult learners of French based on the criterion of lack of a salient foreign accent, but the results showed that the overall grammaticality judgement scores of these participants was distinctively below that of native speakers. However, there is some evidence that reaching nativelike proficiency is possible for late L2 learners. In terms of phonology, findings have shown that some participants were able to pass as native speakers according to the judgement of a panel of native speaker judges (Bongaerts et al., 1995; Bongaerts, van Summeren, Planken, & Schils, 1997; Bongaerts, 1999; Bongaerts, Mennen, & Slik, 2000); or to score within the range of native speakers in pronunciation tasks (Moyer, 1999).

In the area of morphosyntax, there is further evidence that lends support to the possibility of late L2 learners reaching nativelike proficiency. Two exceptionally successful late L2 learners of Egyptian Arabic named Julie and Laura were reported in a case study (Ioup et al., 1994). Julie had not received formal L2 instruction and immigrated to Cairo at the age of 21. Her acquisition of L2 was purely naturalistic, and at the time of participation in the research she had been living in the L2 context for 26 years. In comparison, Laura had the experience of both instructed and naturalistic learning (the exact age of immersion of Laura was not reported in Ioup et

al., 1994). Prior to her immigration to the L2 context, Laura had taken Arabic courses at different universities. At the time of study, Laura had been living in Cairo for 10 years. These two participants were asked to complete a large set of tests that were linguistically demanding, such as a speech production test, two accent identification tests and an anaphora interpretation test. The findings of Ioup et al. (1994) showed that these two participants both performed at the level of native speakers. They concluded that these two successful late L2 learners are exceptions to the critical period hypothesis, and that the influence of the maturation effect had not been a factor for these learners in the way that it is for the majority of L2 learners.

It can be seen that although the possibility of reaching nativelike proficiency for adult learners of an L2 is not as high as for early learners, there are people who actually manage to reach nativelikeness. Depending on the way that nativelikeness is understood, nativelike L2 learners can be categorised into three different types (Abrahamsson & Hyltenstam, 2009). The first type is those who identify themselves as nativelike speakers. Obviously, this kind of self-perceived nativelikeness is not sufficient in academic inquiry. The second type of nativelike speakers is those who are perceived as native speakers by native speakers. The subjects in Bongaerts's (1999) and Moyer's (1999) studies are in this category. Finally, the most rigorous way of conceptualising nativelikeness is linguistic nativelikeness, meaning L2 learners who are like native speakers at a linguistic level. It has been argued that investigations of learners who can pass strict linguistic scrutiny are necessary because it is these learners that provide evidence for or against the hypothesised critical period (e.g. Abrahamsson, 2012; Abrahamsson & Hyltenstam, 2008; 2009; Birdsong, 1999; Hyltenstam, 1992; Hyltenstam & Abrahamsson, 2000; Long, 1990).

It is further argued that if nativelikeness is interpreted as linguistically nativelike, then not a single nativelike adult L2 learner has been found to date. Even the exceptionally successful cases of Julie and Laura diverged from native speakers in some subtle ways (Abrahamsson & Hyltenstam, 2009). This perspective takes linguistic nativelikeness to an extreme and counts L2 learners who deviate very little

from native speakers as examples of failure to reach nativelike proficiency. They suggest that these very successful adult learners are better described as near-native L2 speakers (Abrahamsson & Hyltenstam, 2008; Birdsong, 1999; Hyltenstam & Abrahamsson, 2003). The near-native learners are likely to be perceived as nativelike speakers at first impression, but they may deviate from native norms in linguistically demanding tasks.

2.3.4 Age variables and L2 outcomes

Researchers have empirically investigated different age variables, together with the investigation of other individual difference variables, in order to understand the influence of age on second language acquisition. Different terminologies have been used to refer to the age variable, and the most commonly used are age of acquisition/arrival (AoA), age of first exposure (AoE), and age of onset (AO). AoA is a term usually used in the study of immigrants, and it refers to the age at which learners are immersed in the L2 context. As such, AoA is also often referred to as age of immersion. In contrast, AoE refers to the age that learners are first exposed to L2, and this can occur in a formal classroom setting, or in a visit to the L2 country (Birdsong, 2006). Depending on specific research contexts, AoE and AO could be referring to the same age variable because both of these indicate the starting of second language acquisition (Granena & Long, 2013a). AO is operationalised as the beginning of a serious and sustained process of language acquisition as the result of immigration or the commencement of a formal language program (Granena & Long, 2013a). However, some scholars prefer the term AO rather than AoE because they consider that AO implicates the actual start of the acquisition process (e.g. Abrahamsson & Hyltenstam, 2009; Granena & Long, 2013a; Granena & Long, 2013b).

In the following review of literature, both the terms AoA and AO will be used. In line with the literature, AoA is linked to the acquisition of an L2 in naturalistic

contexts and is used to specifically refer to the age that immersion in the L2 context begins. AO is linked to the learning of an L2 in instructed contexts, which refers to the age at which L2 instruction is first given.

Much previous research that examined the relationship between age and ultimate attainment was conducted in naturalistic contexts. This means that the age variable under investigation in this strand of research is age of immersion (i.e. AoA). In some studies, data of AO was collected, and the relevance of this age variable was examined and compared to age of immersion (e.g. Birdsong & Molis, 2001; Johnson & Newport, 1989). Another variable that is closely related to age, namely length of residence (LoR) in the L2 country, was also examined in relation to L2 outcomes. It is generally suggested that AoA is a more robust predictor than AO and LoR.

L2 outcomes, another important variable in this area of research, have been measured in different ways. For example, L2 learners may be asked to complete grammaticality judgement tests, or to perform a language production task. Tests of L2 outcomes can be categorised into different categories, based on the linguistic area under investigation. For example, phonology and morphosyntax are the two domains that have attracted much empirical attention. Since the present study focuses on the domain of morphosyntax, previous research that targets the same area is reviewed below.

Research on morphosyntax has widely used grammaticality judgement tests (GJTs), both in aural and written modes. This is probably because Johnson and Newport (1989) used such tests, and researchers have followed them in subsequent studies. The results of these studies reveal a significant and rather strong negative correlation between age of immersion and GJT scores, as is shown by the coefficients of -.77 by Johnson and Newport (1989); -.68 by Jia (1998); -.63 by DeKeyser (2000); and -.61 by McDonald (2000), to name a few. In a synthesis of research findings, DeKeyser and Larson-Hall (2005) point out that correlations between age of immersion and written GJT scores are lower than those with aural GJT scores. For example, in Jia's (1998) study, the correlation between AoA and

aural GJT scores was -.68 but was only -.35 with written GJT scores. It is possible that the type of knowledge that is measured by aural and written GJTs is different in nature (Kim & Nam, 2017), so that the age variable appears to have different correlation coefficients with the scores of a GJT of a different modality. The presentation of aural stimuli in GJTs requires automatic processing, which probably draws more on implicit knowledge. In comparison, in written GJTs, the presentation of written stimuli probably provides a chance for heavier reliance on explicit knowledge. Therefore, it could be assumed that age of immersion is more related to implicit L2 knowledge than to explicit knowledge.

In studies that measured L2 outcomes by other methods, such as oral or written production tasks (e.g. Hyltenstam, 1992; Patkowski, 1990), negative correlations between age of immersion and morphosyntax scores were again observed. Production data is more likely a reflection of implicit knowledge than explicit knowledge (R. Ellis, et al., 2009), and thus the negative correlations again indicate the influence of age of immersion on implicit knowledge.

With regard to the role of AO on L2 outcomes, only a few studies have addressed this issue. In all of these studies, the main focus was on the relationship between AoA and ultimate attainment. These studies also show that when AoA and AO are compared, AoA is more influential than AO. For example, Johnson and Newport (1989) collected data of AO alongside AoA, and they found that AO was not significantly correlated with GJT scores (r = -.33, p > .05). Similarly, Birdsong and Molis (2001) reported a nonsignificant correlation between AO and GJT scores. Both of these two studies used Johnson and Newport's (1989) aural GJT as the measure of L2 outcomes. In this test, stimuli were repeated twice to the participants with a 1 to 2 second pause in between. Arguably, such an administration format may have allowed for the use of more explicit knowledge than implicit knowledge because participants probably had enough time to analyse the grammatical correctness of the test stimuli. As suggested by R. Ellis et al., (2009), implicit knowledge is likely to be accessed under time pressure, and tests administered

without time pressure potentially leave a chance for the use of explicit knowledge. Following this argument, it can be assumed that both Johnson and Newport (1989) and Birdsong and Molis (2001) measured explicit knowledge, and therefore the lack of correlation between AO and their GJT scores suggests that AO is not related to explicit L2 knowledge. AO is probably not related to implicit knowledge either because research has shown that AoA is a stronger indicator of L2 outcomes.

Research on the role of LoR on L2 outcomes seems to have produced mixed findings. In research investigating morphosyntax, the recurring pattern is that there is no correlation between LoR and ultimate attainment (Birdsong & Molis, 2001; Johnson & Newport, 1989), and DeKeyser (2000) even found a correlation of exactly zero. On the other hand, there is some evidence that LoR is related to L2 phonology (Flege, Takagi, & Mann, 1995; Flege & Liu, 2001; Purcell & Suter, 1980; Suter, 1976). As discussed above, previous research on morphosyntax is more likely to provide data about participants' explicit knowledge; therefore the lack of correlation between LoR and morphosyntax scores seems to indicate a lack of connection between LoR and explicit knowledge. In comparison, one's knowledge of phonology is probably more implicit than explicit, and the existence of a correlation between LoR and L2 phonology appears to be indicative of a relationship between LoR and implicit knowledge.

One possible reason for these mixed findings of LoR is probably that researchers have set different requirements for minimum LoR in different studies. For example, Johnson and Newport (1989) used a cut-off of 5 years, but DeKeyser (2000) and DeKeyser et al. (2010) set a cut-off of 10 years. It is probably difficult to reach a consensus on how much time would be necessary for L2 learners to fully develop to their ultimate level of attainment, but it is likely that 5 years of residence in the L2 country is not enough. It is possible that even 10 years of residence is not enough either, even if we assume that adults have access to both the implicit and explicit learning systems. The longer one stays in the L2 context, the more one actively uses the L2 in communication, and the higher probability that one's L2

knowledge would be developed. However, it can be argued that after a fairly long period of residence in the L2 country (e.g. 10 years), one's proficiency has reached a relatively stable state (Long, 2003).

2.3.5 Research question two

It is evident from these findings that different age variables play different roles in second language acquisition. Age of immersion (i.e. AoA) is more influential than onset age of learning (i.e. AO), judging from its significant negative correlations with L2 outcomes. However, as the correlations between AoA and tests that potentially access implicit knowledge are higher than those with tests that tap into explicit knowledge, it can be assumed that AoA is more connected to implicit knowledge than explicit knowledge. In comparison, AO is probably related to neither implicit nor explicit knowledge because this variable is shown to have non-significant correlations with L2 outcomes. In the area of morphosyntax, LoR is probably not related to explicit knowledge because in previous studies that employed tests of explicit morphosyntactic knowledge, LoR appears as an irrelevant factor to L2 morphosyntax. However, as LoR seems to play a role for phonology, which can probably be characterised as primarily implicit, it may also play a role for implicit morphosyntactic knowledge. These assumptions need to be tested with the use of separate tests of implicit and explicit morphosyntactic knowledge.

Based on the above summary, the second research question is proposed. This research question asks about the relationship between AO, AoA, LoR and implicit and explicit knowledge in the area of L2 morphosyntax. Accordingly, the following hypotheses are made:

Hypothesis 2a: AO will not influence the implicit knowledge of the NNS.

Hypothesis 2b: AoA will have a negative influence on the implicit knowledge of the NNS.

Hypothesis 2c: LoR will have a positive influence on the implicit knowledge of

the NNS.

Hypothesis 2d: AO will not influence the explicit knowledge of the NNS.

Hypothesis 2e: AoA will not influence the explicit knowledge of the NNS.

Hypothesis 2f: LoR will not influence the explicit knowledge of the NNS.

2.4 Language aptitude and second language acquisition

2.4.1 The construct of language aptitude

Apart from the maturational factor of age, language aptitude has been identified as one relatively robust factor that consistently predicts second language acquisition success. The correlations between aptitude and L2 achievement range mostly between .20 and .60 (Dörnyei & Skehan, 2003), which indicate a rather close relationship between language aptitude and second language acquisition outcomes.

Language aptitude is usually conceptualized as the cognitive abilities that individual learners have for language acquisition. It is a type of mental ability, which means that it refers to the human traits that are being activated when thinking, reasoning, processing information and acquiring new knowledge (Dörnyei, 2006). It is believed that individuals exhibit considerable variation in this kind of ability, which in turn contributes to variation in the outcomes of second language acquisition.

There is considerable debate with regard to the exact nature of language aptitude. Some scholars maintain that language aptitude is an innate trait that is impervious to external influence (Carroll, 1981). According to this view, learner-external factors, such as one's learning experience, do not influence one's aptitude, and aptitude remains stable throughout one's life span. However, empirical studies that explicitly investigate whether or not language aptitude can be influenced by other factors have been limited in number. Only a handful of studies have been conducted to date to examine the extent to which language aptitude is subject to the

influence of learner-external factors, such as learning experience, and learner-internal affective and cognitive factors, such as motivation and anxiety (Eisenstein, 1980; Ganschow & Sparks, 1995; Harley & Hart, 1997; Sáfár & Kormos, 2008; Sparks, Ganschow, Artzer, Siebenhar, & Plageman, 1997). Of these studies, Sáfár & Kormos's (2008) study seems to be the most robust study because it made two kinds of comparisons. The first comparison was between participants who differ in terms of learning experience, and it examined the extent to which the higher scores of participants with more experience are influenced by higher aptitude. The second comparison was made between two aptitude scores of the same group of participants, with the purpose of determining whether improvement in aptitude can be attributed to the amount of learning experience (Li, 2016). The findings of Sáfár and Kormos (2008) suggest that language aptitude can be influenced by external factors, such as intensive language learning experience, which seems to be contrary to the static view of language aptitude.

Rather than viewing language aptitude as a static innate trait, Robinson (2005) took a dynamic perspective and defined aptitude as "cognitive abilities information processing draws on during L2 learning and performance in various contexts and at different stages" (p. 46). This view differs substantially from the perspective of Carroll (1981), because it argues that language aptitude is sensitive to external factors such as learning contexts (Cronbach & Snow, 1977). In turn, this means that there is an interaction between learners' aptitude and teachers' instruction, which would suggest that instruction be adjusted or modified in order to align with learners' level of aptitude.

Although some scholars view language aptitude as a static trait and others contend that it is a dynamic construct, it is generally agreed that language aptitude is a set of cognitive abilities that is distinct from other individual difference variables (e.g. motivation and attitude). Depending on the purposes of a specific research investigation, it seems that both views are plausible. If the primary focus of the research is concerned with the outcomes of second language acquisition, then it

would seem logical to adopt a static view because it denotes a relationship between language aptitude and ultimate L2 attainment. In comparison, if the aim of the research is to understand second language acquisition processes, then it is more appropriate to examine whether language aptitude is malleable in response to contextual factors.

2.4.2 Components of language aptitude

Early theorising about language aptitude was linked to the development of language aptitude tests. These include the Modern Language Aptitude Test (MLAT) developed by Carroll and Sapon (1959), and the Pimsleur Language Aptitude Battery (PLAB) developed by Pimsleur (1966). The primary aim of devising language aptitude tests was to identify individuals who have the potential of learning an additional language within a fairly short period of time. According to Carroll and Sapon (1959), aptitude comprises the following four components: phonetic coding ability, grammatical sensitivity, inductive language learning ability, and rote learning ability (Carroll, 1981). Phonetic coding ability refers to the ability to code unfamiliar sound, so that it can be retained and then be retrieved or recognized. Grammatical sensitivity refers to the capacity to identify the grammatical functions that words fulfil in sentences. Inductive language learning ability refers to the capacity to extract patterns from language materials and to extrapolate from the patterns to produce new sentences. Rote learning ability refers to the ability to form associative bonds in memory between L1 and L2 vocabulary (Dörnyei & Skehan, 2003). Pimsleur's view (1966) was similar; he stated that language aptitude is made up of verbal intelligence, motivation, and auditory ability. Despite the differences in taxonomy, the components of aptitude proposed by Carroll and Sapon (1959) and Pimsleur (1966) overlap to some extent. Pimsleur's verbal intelligence is similar to grammatical sensitivity and inductive language learning ability, and the notion of auditory ability resembles Carroll's proposition of phonetic coding ability (Dörnyei,

2006).

These early aptitude theories did not lead to subsequent extensive research on language aptitude. Second language researchers somewhat marginalised the study of language aptitude because these initial theories were rooted in the teaching context of audio-lingualism, a teaching approach that was gradually abandoned in second language teaching. The four-component view of language aptitude as proposed by Carroll and Sapon (1956) was also challenged. For example, Krashen (1981) maintained that the only component of language aptitude should be grammatical sensitivity, and that phonetic coding and rote memory were irrelevant. Krashen (1981) further argued that there is a link between language aptitude and learning context. Specifically, under the premise of the difference between acquisition and learning, language aptitude is thought to be relevant only to learning, not acquisition. As prevailing teaching methods moved towards the use of more communicative approaches, research on language aptitude did not progress to a significant extent.

Later in the 1990s, Skehan (1998) adapted Carroll's model on the basis of a body of empirical evidence and argued that aptitude would be better viewed as consisting of auditory ability (resembles phonetic coding ability), memory ability (resembles rote memory ability), and linguistic ability (a combination of grammatical sensitivity and inductive language learning ability). Skehan's (1998) tripartite proposal linked language aptitude to different stages of information processing. Auditory ability is thought to be related to input processing, language analytic ability is related to central processing, and memory ability relates to language output and fluency. This is a rather parsimonious and general view of language aptitude, which is consistent with a cognitive view of second language acquisition (Dörnyei & Skehan, 2003). Skehan's (1998) initial proposal was later extended in more detail, covering more refined language acquisition stages such as noticing, pattern identification, and pattern integration (Skehan, 2002; 2012). Other aptitude components, such as working memory, were also proposed in accordance with its potential relevance to putative second language acquisition stages.

Working memory

Although Carroll's MLAT does have a memory component, Carroll (1990) was not confident about the validity of this subset. However, Carroll's view of the memory component in language aptitude was limited to the passive storage capacity of memory. In comparison, some other researchers found that working memory, which is slightly different from rote memory, is very important to language acquisition (Baddeley, 1992; 1998; 2007; Baddeley & Hitch, 1974). Working memory is shown to be different from long-term and short-term memory, which in Skehan's (1998; 2002; 2012; 2015b) view is implicated at the initial stages of L2 acquisition, including input processing, noticing and pattern identification.

As its name indicates, working memory is responsible for both temporarily storing and processing information (Miyake & Friedman, 1998). It can be defined as the "set of processes that hold a limited amount of information in a readily accessible state for use in an active task" (Cowan, 2005, p. 39), and in which multiple cognitive processes including phonemic coding, short-term memory, and language processing are involved (Baddeley, 1999; Li, 2013). Three components of working memory have been proposed, namely the central executive, the phonological loop and the visuo-spatial sketch pad (Baddeley, 1999). The central executive is the main component, which is limited in capacity and is used for both information storage and processing. The phonological loop and the visuo-spatial sketch pad are two slave systems that serve the central executive, with the former responsible for the storage and rehearsal of verbal information and the latter responsible for the retention of visual information.

Some scholars maintain that it is in working memory that the three components of language aptitude (phonetic coding, language analytic ability and memory) converge (Miyake & Friedman, 1998), and a number of scholars argue forcefully that working memory constitutes a component of language aptitude (DeKeyser & Koeth, 2011; Sawyer & Ranta, 2001; Skehan, 2012; 2015b; Wen, 2014; 2015; Wen & Skehan, 2011). This view of the relationship between working memory and

language aptitude is also reflected empirically, as tests of working memory have been incorporated into one recently devised language aptitude test battery, namely the Hi-LAB (Linck et al., 2013). In this test battery, six measures of working memory targeting both the storage and the executive functions are included. Empirically, a distinction between phonological short-term memory and executive working memory is generally made, and it has been suggested that these two working memory components are directly relevant to language acquisition (Baddeley & Gathercole, 1992; Juffs & Harrington, 2011; Linck, Osthus, Koeth, & Bunting, 2014; Wen, 2012; 2014; Williams, 2012). Phonological short-term memory is thought to be responsible for the storage function, whereas executive working memory performs both the storage and the processing functions.

In a recent meta-analysis (Li, 2016), the relationship between working memory and language aptitude was examined based on a synthesis of research findings to date. It was found that executive working memory had moderate but significant correlations with language aptitude (as a holistic construct) (r = .33) and with phonetic coding ability (r = .33). The correlations between executive working memory and language analytic ability and rote memory were significant, but were weaker (r = .22 and .25, respectively). On the other hand, phonological short-term memory was found to have yet weaker or nonsignificant correlations with language aptitude and its sub-components. Based on these findings, Li (2016) suggests that executive working memory is potentially a more promising language aptitude component than phonological short-term memory. However, Li (2016) also calls for further research because there is a possibility that these two kinds of working memory abilities are implemented at different stages of language acquisition. It is also possible that these two working memory abilities are related to different aspects of second language acquisition. For example, phonological short-term memory has been suggested as an important factor that influences vocabulary learning (Baddeley, Gathercole, & Papagno, 1998), whereas executive working memory plays a role in reading and listening comprehension due to its dual function of storage and

processing (Daneman & Merikle, 1996). Nonetheless, despite these areas of uncertainty, researchers have called for the inclusion of working memory as a component of language aptitude at a theoretical level (DeKeyser & Koeth, 2011; Robinson, 2002b; Skehan, 2012).

Language analytic ability

Language analytic ability is considered to be another important aspect of language aptitude. This ability combines the components of grammatical sensitivity and inductive learning ability, and represents an analysis-oriented ability in general (Skehan, 1998). It can be defined as "the capacity to infer rules of language and make linguistic generalizations and extrapolations" (Skehan, 1998, p. 207).

Many studies have provided evidence of the robustness of language analytic ability in language learning (Abrahamsson & Hyltenstam, 2008; Bylund, Abrahamsson, & Hyltenstam, 2009; Dörnyei, 2010; Granena & Long, 2013a; Granena, 2014; Harley & Hart, 2002; Robinson, 2001). This component has been considered as representative of language aptitude by some researchers because of its central role (e.g. DeKeyser, 2000; DeKeyser et al., 2010). This ability has been particularly associated with language acquisition in contexts where there is more of an emphasis on explicit instruction (Krashen, 1981). Supporting evidence comes from research on intensive language training programs (Ehrman & Oxford, 1995), immersion programs (Harley & Hart, 1997), and also the acquisition of artificial languages (de Graaff, 1997). For example, in a survey-based study (Ehrman & Oxford, 1995), language analytic ability was found to be the most salient factor indicating overall learning success after intensive language training. As it is assumed that conscious learning is most likely to be involved in this kind of intensive training context, it is therefore concluded that language analytic ability could be very relevant to explicit learning processes. However, as DeKeyser (2000) points out, the "conscious-unconscious dichotomy does not completely coincide with the instructed-naturalistic distinction" (p. 507). This means that even in naturalistic

learning contexts, a certain degree of consciousness could be involved and used as a tool to reflect on the structure being acquired. If this is the case, it seems that there is a possibility that language analytic ability is also implemented in naturalistic learning contexts. In other words, it can be assumed that language analytic ability might be relevant to both implicit and explicit learning processes.

2.4.3 Measurement of language aptitude

The earliest language aptitude test batteries that were developed in the 1950s and 1960s, the MLAT (Carroll & Sapon, 1959) and the PLAB (Pimsleur, 1966) are the most frequently used aptitude measures to date. The popularity of these two test batteries is largely a result of their strong predictive validity, which was established based on validation with a large number of participants (5000 learners for the MLAT and 6000 for the PLAB). The MLAT was even regarded as "the best overall instrument for predicting language learning success" (Parry & Child, 1990, p. 52). In the following decades, a number of other language aptitude test batteries have been developed, these include the Defense Language Aptitude Battery (DLAB) (Petersen & Al-Haik, 1976); the VORD (Child, 1998); the Cognitive Ability for Novelty in Acquisition of Language (Foreign) (CANAL-F) (Grigornko, Sternberg, & Ehrman, 2000); the LLAMA test (Meara, 2005); and the Hi-LAB (Linck et al., 2013).

Due to discrepancies in the conceptualisation of the construct of language aptitude, there are considerable differences in the way that language aptitude is defined and operationalised in different aptitude test batteries. For example, the Hi-LAB (Linck et al., 2013) emphasized working memory ability, a component that was not measured by other aptitude tests such as the MLAT and the PLAB. Other test batteries were largely based on Carroll's framework of language aptitude, which means that the kind of abilities that these test batteries measure were similar. The CANAL-F test battery (Grigornko et al., 2000) took a slightly different theoretical perspective from the other tests, in the sense that it was rooted in a particular

cognitive theory of foreign language acquisition, which emphasizes the ability to deal with linguistic materials in learning contexts that mimic instructed language learning settings. However, this test appears to perform very similarly to previous tests, albeit being grounded on a different theoretical perspective (R. Ellis, 2008b).

Since measures of working memory are generally not included in language aptitude test batteries, researchers have been using individual tests to measure working memory ability. As stated earlier, the empirical distinction between short-term working memory and complex working memory has generally been made. Short-term working memory refers to phonological short-term memory that only has a passive storage function, while executive working memory refers to the ability to store and process information. Common measures used to measure short-term memory are nonword repetition tasks, digit span tasks, and reverse digit span tasks; while reading or listening sentence span tasks are used to measure executive working memory (Skehan, 2010; Wen, 2015).

The one test battery of language aptitude that has gained in popularity in recent years is the LLAMA test (Meara, 2005). This test is modelled on the MLAT (Carroll & Sapon, 1959), but it has several other merits which have contributed to its increasing implementation in the testing of language aptitude. Firstly, this test is language neutral, which means that it can be used with different learner populations. This language neutral characteristic is realised by including picture stimuli, so that the possibility of L1 interference in test completion is avoided (see Section 3.4.5 for detailed description of this test). Secondly, the sub-test LLAMA D has been suggested as a measure of implicit language aptitude (Granena, 2012; 2013a), which is an improvement on previous aptitude test batteries because it potentially accesses a kind of ability that is implicated in naturalistic learning conditions. Last but not least, the accessibility of the LLAMA test is better than that of previous tests such as the MLAT, and therefore this test has inevitably attracted the attention of researchers.

However, it should be noted that the LLAMA test has only been validated very recently (Rogers, Meara, Barnett-Legh, Curry, & Davie, 2017). At the initial stages

of the present study, the only studies that had examined the constructs that this test measures were Granena (2012; 2013a). Based on her results of a principle component analysis (PCA), Granena (2012; 2013a) presented some evidence that the LLAMA test was likely a measure of two different components, namely sequence learning ability as measured by LLAMA D and language analytical ability as measured by a combination of LLAMA B, E and F. She further argued that LLAMA D measured an implicit aptitude component, while the other three sub-tests measured an explicit aptitude component. Rogers et al. (2017) presented a more thorough validation of the LLAMA test, based on data collected from 240 participants of different ages. Contrary to Granena's (2012; 2013a) findings, the results of Rogers et al. (2017) suggested that LLAMA D was not a likely measure of implicit aptitude. The reason was that if it measured some implicit ability, then learners at younger ages should have demonstrated an advantage in the scores of this sub-test in comparison with older learners. However, they found that the younger participant group (aged 10 to 11 years) performed worse than the other two groups (aged 20 to 21 and 30 to 70, respectively) in LLAMA D. In subsequent regression analysis, Rogers et al. (2017) further examined the extent to which individual difference factors including gender, age and education levels, influence LLAMA scores, and they concluded that these factors do not affect overall LLAMA test performance. Clearly, what the LLAMA test actually measures awaits further investigation, as Rogers et al. (2017) noted "we make no claims regarding how well they (i.e. the four sub-tests of the LLAMA test) measure aptitude (however defined)" (p. 56).

2.4.4 Language aptitude, acquisition contexts and L2 outcomes

Since the introduction of the MLAT, there has been considerable focus on developing language aptitude test batteries. However, some scholars have taken another approach and have endeavoured to investigate the extent to which language aptitude operates in different learning contexts. For example, Reves (1983), as cited

in Skehan (2002), concludes that language aptitude plays a role in both naturalistic and instructed learning contexts. This study examined the predictive power of language aptitude on acquisition outcomes by comparing a group of Arabic L1 learners of Hebrew in a naturalistic context and Arabic L1 learners of English in an instructed context. It was found that language aptitude functioned as an effective predictor of acquisition outcomes in both contexts, which contradicted Krashen's (1981) hypothesis that language aptitude is only relevant to learning.

Similar to Reves's (1983) conclusion, Skehan (1989) predicted that language aptitude should be relevant in naturalistic learning contexts because learners would encounter a higher level of difficulty in extracting structures from linguistic input than in instructed contexts. Without explicit rule-based instruction, the duty of processing massive amounts of input in naturalistic contexts lies exclusively on learners themselves (McLaughlin, 1990). Therefore, if learners' individual differences have an impact on learning outcomes in classroom settings in which learning materials are selected and sequenced, they should have even more of an impact on acquisition outcomes in naturalistic contexts. In particular, language analytic ability, representing one's ability to handle language structures in general, has been argued as an important language aptitude component that figures in naturalistic learning contexts (Skehan, 1986; 1989; 1998). This view gains support from research in aptitude-treatment interaction, which shows that the heavier the information processing burden that is put on students, the more important the role of language aptitude (e.g. Snow, 1991). Further evidence comes from studies that adopted an experimental design. For example, Robinson (1995b) asked participants to try to understand the sentences that were given to them, which created a learning condition that was similar to naturalistic learning contexts. A more explicit learning condition was also created in which learners were asked to look for rules (i.e. an inductive condition) or to learn the rules as they were explicitly taught (i.e. a deductive condition). The role of language aptitude in these different learning contexts was then examined, and it was found that language aptitude was implicated

in both implicit and explicit learning contexts.

On the other hand, there is general consensus that language aptitude figures in instructed learning contexts. The first language aptitude test batteries (i.e. the MLAT and the PLAB) were designed for learners in instructed contexts, and their validity for predicting learning outcomes has been demonstrated with a large number of learners (Carroll & Sapon, 1959; Pimsleur, 1966). Further evidence that lends support to the robustness of language aptitude in instructed contexts comes from experimental studies that involve classroom learners who were treated with implicit or explicit instructional methods (Alderson, Clapham, & Steel, 1997; de Graaff, 1997; Erlam, 2005; Hwu & Sun, 2012; VanPatten & Borst, 2012a; 2012b). Findings of these studies demonstrate that learners who have higher aptitude tend to benefit more from instruction, including both implicit and explicit instruction (Skehan, 2015a). Therefore, it has been suggested that language aptitude is a conscious construct that affects learning outcomes in explicit learning conditions (Li, 2015). However, it should be noted that these experimental studies were conducted within a rather short period of time, lasting for several weeks or a semester. This means that the learning outcomes being measured in these studies were more relevant to learning rate, rather than long-term language attainment.

Working memory ability, as a proposed aptitude component, has been suggested as an important factor that influences second language learning outcomes. Previous research findings have shown a close relationship between phonological short-term memory and the development of various linguistic aspects, including vocabulary (Cheung, 1996; N. C. Ellis, 1996; French, 2006; Service, 1992), formulaic sequences and collocations (Bolibaugh & Foster, 2013; Foster, Bolibaugh, & Kotula, 2014; Skrzypek, 2009), morpohosyntax and grammar (Martin & Ellis, 2012; O'brien, Segalowitz, Collentine, & Freed, 2006; O'brien, Segalowitz, Freed, & Collentine, 2007; Williams & Lovatt, 2003), and L2 comprehension and production(O'brien et al., 2006; 2007; Payne & Whitney, 2002). Similarly, there has been some evidence suggesting that executive working memory is relevant to L2 comprehension and

production (Alptekin & Erçetin, 2011; Harrington & Sawyer, 1992; Walter, 2004). Based on these findings, it is therefore concluded that working memory, as comprised of phonological short-term memory and executive working memory, is related to L2 outcomes.

In spite of this considerable research evidence, the extent to which language aptitude influences learning outcomes, where outcomes are understood as implicit and explicit knowledge, awaits further investigation (Li, 2016). This is mainly a result of different views of aptitude components and the empirical difficulties in assessing implicit and explicit knowledge. As reviewed above, language analytic ability and working memory capacity are both potentially influential on L2 outcomes. However, it is probably very difficult to make inferences about the role of individual components as there is a lack of validation studies of the language aptitude measures, especially for the LLAMA test. What remains to be established is the following: 1) Is the LLAMA test a valid measure of language aptitude?, 2) Is working memory ability a language aptitude component?, and 3) How do language aptitude and working memory relate to L2 outcomes, understood as implicit and explicit knowledge?

2.4.5 Research question three

The third research question is addressed in the light of previous research. This question asks what the relationship is between language aptitude, working memory and implicit and explicit L2 knowledge. Accordingly, three hypotheses are made as follows:

Hypothesis 3a: Language aptitude will have a positive influence on the implicit knowledge of the NNS. This influence is mainly contributed by sequence learning ability as measured by LLAMA D.

Hypothesis 3b: Language aptitude will also have a positive influence on the explicit knowledge of the NNS. This influence is contributed by language analytic

ability as measured by a combination of LLAMA B, E and F.

Hypothesis 3c: Working memory will have a positive influence on both the implicit and the explicit knowledge of the NNS.

2.5 The interaction of age, aptitude, and L2 knowledge

2.5.1 Empirical evidence

Given the fact that both aptitude and age have been treated as robust predictors in second language acquisition research (e.g. Erlam, 2005; Flege et al., 1999; Granena & Long, 2013a; Granena, 2014; Harley & Hart, 1997; 2002; Jia & Fuse, 2007; Patkowski, 1990), it is natural to wonder if there is an interaction between these two variables. It can be assumed that, based on current findings, the age-aptitude covariant may result in diverse ultimate levels of proficiency, where proficiency is operationalised as implicit and explicit L2 knowledge. To the best of the author's knowledge, five studies have been conducted to examine the age, aptitude and ultimate attainment relationship within the domain of morphosyntax (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; DeKeyser et al., 2010; Granena & Long, 2013a; Granena, 2014).

DeKeyser (2000) and DeKeyser et al. (2010) formed a hypothesis that addresses the issue of age-aptitude interaction in relation to the implicit and explicit distinction of second language acquisition. His hypothesis formulation is based on the fundamental difference hypothesis (Bley-Vroman, 1989), and the main hypotheses can be summarised as follows:

1) Age only influences the capacity of implicit learning. As learners grow older, the implicit learning system is no longer available, so the explicit learning system is used to acquire L2.
2) Language aptitude, particularly language analytic ability, can compensate for learners' declining implicit learning ability. Therefore, it is possible for

adult learners to reach a high level of L2 proficiency provided that they have a high level of language aptitude.

3) The negative influence of age on ultimate L2 attainment is mediated by aptitude.

The above listed five studies provide some empirical evidence in relation to these hypotheses, but the findings have been mixed. DeKeyser (2000) and DeKeyser et al. (2010) found that there were significant correlations between aptitude and ultimate attainment in grammatical L2 features with older learners, but not with younger learners, while Abrahamsson and Hyltenstam (2008) found that aptitude-ultimate attainment correlations in both early and late learners. Granena and Long (2013a) empirically tested the possibility of multiple sensitive periods, and their findings lent support to their hypotheses. In pronunciation, aptitude was found to be relevant for both early and late learners, while in lexis and collocation, aptitude was only relevant for late learners. Similarly, Granena (2014) found that aptitude is relevant for early learners in an untimed GJT, but the relevance of aptitude may be restricted to explicit learning contexts rather than implicit learning contexts. In the following paragraphs, these five studies are reviewed in more detail.

All of these studies examine acquisition outcomes in the morphosyntactic domain, except for Granena and Long's study (2013a), which also investigates the phonological, lexical and collocational aspects of language. It can be seen that these studies accept the maturational constraints account of second language acquisition, as they either acknowledge the existence of a critical/sensitive period (e.g. DeKeyser, 2000) or propose different critical/sensitive period(s) for different linguistic domains (e.g. Granena & Long, 2013a). However, there are considerable differences in many aspects of these studies, including participant sampling strategy, research contexts, and measures used.

Following different participant sampling strategies, different learner populations have been involved in these studies. DeKeyser (2000) and DeKeyser et al. (2010) used advertisements as the main approach to recruit participants. DeKeyser's (2000)

study involved 57 native speakers of Hungarian, who had been learning English in the US. Among these participants, 42 had an AoA of more than 16 years old, and the other 15 immigrated to the US before the age of 16 years. All of them had more than 10 years of LoR, with the mean LoR for the younger group 35.6 years, and 33.7 years for the older group. The average age at the time of study was 55 years old. In DeKeyser et al.'s (2010) study, two groups of ESL learners were involved. There were a group of 76 native speakers of Russian, who had learnt English as an L2 in the US. The minimum LoR of this group was 8 years, the AoA ranged from 5 to 71 years old, and age at testing ranged from 19 to 79 years old. The other group, 62 native speakers of Russian acquiring Hebrew as an L2, also had a minimum LoR of 8 years, with AoA ranging from 4 to 65 years old, and age at testing ranging from 16 to 75 years old.

In contrast, the remaining three studies involved some kind of screening process in selecting potential participants. In particular, Abrahamsson and Hyltenstam's (2008) screening process was quite stringent. First, a total number of 195 advanced L2 speakers of Swedish with Spanish as L1 were obtained through advertisements and posters. They were then asked to participate in a 15 minutes interview and provide a sample of spontaneous speech on a given topic for 1 minute. The interview and speech samples were recorded, and the recordings of an additional 20 native speakers were obtained in a similar fashion. These samples were then judged by 10 linguistically untrained judges to indicate whether they were produced by native Swedish speakers or not. After this, 104 participants were judged to be perceivably nativelike, and finally 42 individuals were selected from these 104 according to their basic background criteria. The LoR of these participants ranged between 12 to 42 years, and their AoA ranged between 13 to 23 years. Comparatively speaking, the screening process of Granena and Long's (2013a) and Granena's (2014) studies was much simpler: via telephone interviews. Sixty-five native speakers of Chinese who spoke Spanish as L2 participated in Granena and Long's (2013a) study. The LoR of these participants ranged from 8 to 31 years, with an AoA ranged between 3 to 29

years old. Participants in Granena's (2014) study were all early sequential Chinese L1- Spanish L2 bilinguals. Their LoR ranged from 11 to 28 years, and AoA ranged between 3 to 6 years old. It can be seen that the learner populations differed in previous studies, and this difference might to some extent lead to different research findings.

Besides the differences in learner populations involved, a number of different language aptitude measures were used in these studies. Depending on how the construct of language aptitude was operationalised, some studies measured one aptitude component, namely language analytic ability (DeKeyser, 2000; DeKeyser et al., 2010), while others adopted more comprehensive language aptitude measures (Abrahamsson & Hyltenstam, 2008; Granena & Long, 2013a; Granena, 2014). DeKeyser (2000) used a Hungarian Language Aptitude Test, which was an adaption of the Words in Sentences part of the MLAT. DeKeyser et al. (2010) used a Russian version of the Inter-University Psychometric Entrance Test, which was thought to be comparable to the verbal Scholastic Aptitude Test in the US. In comparison, the Swansea Language Aptitude Test was used by Abrahamsson and Hyltenstam (2008), and its latest version, the LLAMA test (Meara, 2005), was used by Granena and Long (2013a) and Granena (2014). As reviewed earlier, although some evidence suggests that the LLAMA test tapped two different aspects of language aptitude (i.e. language analytic ability and sequence learning ability) (Granena, 2012; 2013a), what this test exactly measures remains to be investigated. Nevertheless, suffice to say that language aptitude was operationalised as different constructs by these studies.

In terms of learning outcomes, it is relatively certain that all of these studies did measure ultimate L2 attainment since the average LoR was more than 8 years (Long, 2003). A GJT was used in examining the morphosyntactic proficiency in all of the five studies. DeKeyser (2000) and DeKeyser et al. (2010) employed an unspeeded oral GJT, while Abrahamsson and Hyltenstam (2008) used both a written and an auditory GJT with a time pressure of 10 seconds for each stimulus. Granena (2014)

used both a speeded auditory GJT (response time of each stimulus varied accordingly to pilot study) and a non-speeded one in her study. In Granena and Long (2013a), five different measures were utilized to assess morphosyntactic proficiency, including an auditory GJT, an oral retelling task, two word-order preference tasks, and one gender-assignment task. Measures of phonology, lexis and collocations were particularly designed for the participants, including a read-aloud task and a series of collocation-related judgement tasks.

Given the different measures of aptitude and language proficiency, it is not surprising that research findings are mixed. It is very likely that methodological differences have affected the resultant research findings. DeKeyser's two studies are only concerned with language analytic ability, which is the type of ability that unspeeded GJTs tend to measure. Therefore, it has been pointed out that the correlation between aptitude and morphosyntactic attainment among late starters in DeKeyser's studies is due to this methodological design: both the aptitude measure and the proficiency measure favour similar explicit analytic ability, the ability that is supposed to be better possessed by late starters of an L2 rather than early starters (Granena and Long, 2013a). Compared to DeKeyser's research, Abrahamsson and Hyltenstam's (2008) study employed a more complex and demanding GJT task, and they obtained different findings. Aptitude was found to be relevant to the proficiency scores of both early and late learners, and this led to the conclusion that language aptitude "plays not only a crucial role for adult learners but also a certain role for child learners" (Abrahamsson & Hyltenstam, 2008, p. 499). Granena (2014) obtained similar results to Abrahamsson and Hyltenstam (2008), which showed that aptitude does play a role in early childhood learners' ultimate attainment. In contrast, Granena and Long (2013a) did not find any correlations between aptitude and ultimate morphosyntactic attainment, and this might be a result of the various language measures that were employed (multiple abilities including analytic ability were examined).

One aspect that has been neglected by previous studies is the distinction between implicit and explicit L2 knowledge. This distinction is also relevant to the test modality of GJTs, as R. Ellis (2004, 2005, 2009) pointed out: timed and untimed GJTs tapped into different kinds of linguistic knowledge. Unfortunately, most of the above reviewed studies did not discuss their results in relation to implicit and explicit knowledge (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; DeKeyser et al., 2010; Granena & Long, 2013a), whereas the only one that did so confirmed that timed and untimed GJTs correlate differently with aptitude scores (Granena, 2014).

It is therefore concluded that research of the age-aptitude-ultimate L2 attainment relationship is still very limited to date. Aptitude measures that involve a wider range of components in addition to analytic ability, such as working memory, would be a positive addition to the research in this area. Also, based on this review of the literature, the inclusion of measures of both implicit and explicit knowledge would help to clarify the age-aptitude-ultimate attainment relationship, and also to reconcile previous conflicting research findings.

2.5.2 Research question four

Given the extant research findings, the last research question is posed. This question asks two sub questions: 1) What is the joint influence of age and language aptitude on implicit and explicit L2 knowledge? 2) Is there an interaction between age and aptitude? Accordingly, the following three hypotheses are made:

Hypothesis 4a: Both age and language aptitude will influence the implicit knowledge of the NNS.

Hypothesis 4b: Language aptitude will influence the explicit knowledge of the NNS, but age will not.

Hypothesis 4c: There is an interaction between age and language aptitude. Age will negatively influence the implicit knowledge of the NNS, but this influence will

be mitigated by high language aptitude.

Relevant literature has been reviewed in this chapter, and research questions have been raised. In the next chapter, research methodology will be addressed.

CHAPTER THREE

METHODOLOGY

In this chapter, the research methodology is presented. A description of the participants will be presented first, followed by an overview of the target structures of this study. Then, research instruments are described in detail. Finally, the reliability of the test instruments is addressed and the results of the pilot study are discussed.

3.1 Participants

In the present study, there were two groups of participants; a group of 20 NS of English, and a group of 86 NNS of English. After potential NS participants expressed interest in the study, the researcher verified that they met the following criteria: 1) they had spoken only English at home during childhood; 2) they had had English as the language of instruction at school; and 3) they had lived their whole life in a context where English has been the majority language (Abrahamsson & Hyltenstam, 2008). These 20 participants were not necessarily monolingual speakers of English.

Following the recommendations of previous research (Andringa, 2014; Long, 2003), the present study did not intend to select a highly educated NS group, but it turned out that the NS participants who volunteered held at least a Bachelor's degree or were pursuing one at the time of participation in this study. This might be because the researcher put several advertisements in her university and therefore NS participants who studied at the university or were visitors to the university volunteered. The researcher also put advertisements in two Anglican churches that were close to where she lived, and other NS participants were recruited as a result. Therefore, the NS group was made up of people of different ages, including four university students in their 20s (the youngest) and two middle-aged participants in

their 50s (the oldest ones). The other NS participants were aged around 30 to 40 at the time of participation.

The sample size of the NNS participants was determined based on considerations of the requirements of confirmatory factor analysis. The *N:q* rule was followed, where *N* refers to the number of cases (i.e. participants) and *q* refers to model parameters that require estimates (Jackson, 2003). According to the minimum ratio of 5:1, a structural model of two factors and six indicators requires at least 65 participants (Kline, 2011). However, as larger sample sizes generally benefit factor analysis, the researcher endeavoured to recruit as many participants as possible and finally a total number of 86 NNS participants were recruited.

One of the criteria for the NNS who were interested in participating in the study was that they had been living in New Zealand for more than 8 years. Eighty-six participants were recruited who met this criterion. The minimum LoR requirement was originally set at 10 years to ensure that the NNS had reached a relatively stable stage in their L2 (DeKeyser, 2013; Oyama, 1978). In other words, by asking for a minimum LoR of 10 years (DeKeyser, 2000; DeKeyser et al., 2010), an effort was made to maximise the possibility of investigating the ultimate attainment of the L2 of the NNS, rather than the rate of acquisition. However, at a later stage of data collection, the minimum LoR was reduced to 8 years in an attempt to involve more participants who had been immersed in the L2 context before puberty (≤ 13 years old), and five participants whose LoR was 8 years were recruited. The other 79 NNS had a LoR of more than 10 years.

The other criterion for the NNS group was that they were native speakers of Mandarin Chinese and L2 learners of English. These participants were recruited by advertising among the Chinese community in Auckland, distributing fliers in universities, and word of mouth. By employing such a group of NNS participants who spoke the same L1, the possibility of including the L1 as a variable was avoided (DeKeyser, 2013). A participant was considered as a native speaker of Mandarin if he or she 1) was born to Mandarin speaking parents in China; 2) had spoken only

Mandarin at home before immigration, and 3) had continued to use Mandarin at home/with friends after immigration. Before including them in the study, the researcher made sure that all the NNS participants met all three criteria.

In addition, to ensure that the NNS participants did not live in a primarily Mandarin speaking context after immigration and had had ample opportunities to receive linguistic input and develop their L2 knowledge in the L2 context (DeKeyser, 2013), it was required that in order to be eligible for participation in the study, that all the NNS had to use English as a medium of communication either for work, study, or social purposes. Since all the NNS also had to have lived in an English-speaking country for more than 8 years at the time of participation, it was expected that they had had opportunity to acquire the L2 by using it in the L2 environment.

A questionnaire was used to gather demographic information of the NNS, their experience of learning English in China and New Zealand, and English language use information. This questionnaire is described in detail in Section 3.4.6, and the data gathered from it is presented in Section 4.12.

3.2 Research questions and hypotheses

The present study investigates the influence of age and language aptitude on long-term outcomes of second language acquisition, as examined by a series of tests of implicit and explicit L2 knowledge. Different age variables were investigated, including the AO (age of onset) which refers to the age that the learning of English began, usually in an instructed context, and the AoA (age of arrival) which refers to the age that one started to be immersed in an English context. The other variables that were of research interest were LoR (length of residence) in the L2 country, language aptitude and working memory. To understand the influences of these variables on long-term implicit and explicit L2 knowledge, the following research questions were asked. In relation to the research questions, a series of research hypotheses were made.

RQ1: *What is the profiling of the implicit and explicit knowledge of the NNS and the NS in the present study? Is there a difference between the implicit and explicit knowledge of the NNS in comparison to the NS?*

Hypothesis 1a: The NNS will have a higher level of explicit knowledge than implicit knowledge in English.

Hypothesis 1b: The NS will have a higher level of implicit knowledge than explicit knowledge in English.

Hypothesis 1c: The level of implicit knowledge of the NNS will be lower than that of the NS.

Hypothesis 1d: The level of explicit knowledge of the NNS will be higher than that of the NS.

RQ2: *What is the relationship between age, including AO (age of onset) and AoA (age of arrival); LoR (length of residence), and implicit and explicit L2 knowledge?*

Hypothesis 2a: AO will not influence the implicit knowledge of the NNS.

Hypothesis 2b: AoA will have a negative influence on the implicit knowledge of the NNS.

Hypothesis 2c: LoR will have a positive influence on the implicit knowledge of the NNS.

Hypothesis 2d: AO will not influence the explicit knowledge of the NNS.

Hypothesis 2e: AoA will not influence the explicit knowledge of the NNS.

Hypothesis 2f: LoR will not influence the explicit knowledge of the NNS.

RQ 3: *What is the relationship between language aptitude, working memory and implicit and explicit L2 knowledge?*

Hypothesis 3a: Language aptitude will have a positive influence on the implicit knowledge of the NNS. This influence will be mainly contributed by sequence learning ability as measured by LLAMA D.

Hypothesis 3b: Language aptitude will also have a positive influence the explicit knowledge of the NNS. This influence is contributed by language analytic ability as measured by a combination of LLAMA B, E and F.

Hypothesis 3c: Working memory will have a positive influence on both the implicit and explicit knowledge of the NNS.

RQ4: *What is the joint influence of age and language aptitude on implicit and explicit L2 knowledge? Is there an interaction between age and aptitude?*

Hypothesis 4a: Both age and language aptitude will have an influence on the implicit knowledge of the NNS.

Hypothesis 4b: Language aptitude will influence the explicit knowledge of the NNS, but age will not.

Hypothesis 4c: There is an interaction between age and language aptitude. Age will negatively influence the implicit knowledge of the NNS, but this influence will be mitigated by high language aptitude.

3.3 Target structures

Since the majority of the language tests were adapted from R. Ellis (2005) and R. Ellis et al. (2009), the present study targeted the same 17 linguistic structures as R. Ellis et al. (2009). These structures represent morphological and syntactical aspects of English grammar. As discussed in R. Ellis et al. (2009), these structures were chosen because learners tend to make mistakes in using them. In addition, in terms of the developmental order of L2 acquisition, these structures represent both early and late acquired grammatical structures. Lastly, these selected structures were introduced in ESL courses at different proficiency levels, including beginner, intermediate and advanced levels. Therefore, including a range of structures such as these is suitable for testing participants with a mixed level of proficiency in English. The following table is reproduced from R. Ellis et al. (2009, pp. 43-44), and sets out

the target structures.

Table 3.1 Target structures

Structure	Example of learner error	When acquired	Grammatical type
Verb complements	Liao says he wants *buying* a new car.	Early	Syntactical
Regular past tense	Martin *complete* his assignment yesterday.	Intermediate	Morphological
Question tags	We will leave tomorrow, *isn't it*?	Late	Syntactical
Yes/No questions	Did Keiko *completed* her homework?	Intermediate	Morphological
Modal verbs	I must *to brush* my teeth now	Early	Morphological
Unreal conditionals	If he had been richer, she *will* marry him.	Late	Syntactical
Since and *for*	He *has been living* in New Zealand *since* three years.	Intermediate	Syntactical
Indefinite article	They had *the* very good time at the party.	Late	Morphological
Ergative verbs	Between 1990 and 2000 the population of New Zealand *was increased*.	Late	Syntactical
Possessive -s	Liao is still living in his rich *uncle* house.	Late	Morphological
Plural -s	Martin sold a few old *coin* to a shop.	Early	Morphological
Third person -s	Hiroshi *live* with his friend Koji.	Late	Morphological
Relative clauses	The boat that my father bought *it* has sunk.	Late	Syntactical
Embedded questions	Tom wanted to know what *had I done.*	Late	Syntactical
Dative alternation	The teacher explained *John the answer*.	Late	Syntactical
Comparatives	The building is *more bigger* than your house.	Late	Syntactical
Adverb placement	She writes *very well* English.	Late	Syntactical

3.4 Instruments

3.4.1 Overview

A battery of 12 tests was administered, including 6 language tests of implicit and explicit knowledge, 2 tests of working memory and 4 tests of language aptitude.

In addition, as has already been mentioned, a questionnaire was developed and used by the researcher to ask questions about the English learning experience and current language use situation of the NNS.

Four tests were hypothesised to be measures of implicit knowledge. Among these four, two tests from R. Ellis et al. (2009) were adapted for the present study, including an oral elicited imitation test (EI) developed by Erlam (2006; 2009) and a timed grammaticality judgement test (TGJT) developed by Loewen (2009). The third hypothesised test of implicit knowledge was a word monitoring test (WordM), and this test was developed by the researcher following the suggestions of Suzuki and DeKeyser (2015). The last hypothesised test of implicit knowledge was a cloze test, which was a test originally developed by Davies (1991).

To measure explicit knowledge, two tests of R. Ellis et al. (2009) were employed, including an untimed grammaticality judgement test (UGJT) developed by Loewen (2009) and a metalinguistic knowledge test (MKT) developed by Elder (2009).

Language aptitude was measured using the LLAMA test battery with four sub-tests (Meara, 2005). Two tests of working memory (Zhao, 2015) were used, namely a Mandarin non-word repetition test (NonWord) and a Mandarin listening span test (SSpan).

Table 3.2 provides an overview of these instruments. The NNS participants completed all 12 tests and the questionnaire in a fixed order. They completed the two tests of working memory first (NonWord and SSpan), and then the tests of language aptitude (LLAMA), followed by the EI, the TGJT, the UGJT, the MKT, and the cloze test. Lastly, they filled out the questionnaire. However, the NS participants only completed the tests of implicit and explicit knowledge in the following order: the EI, the TGJT, the UGJT, the MKT and the cloze test. The researcher did not ask the NS participants to fill out a questionnaire to gather information about their age and level of education but interviewed the participants to obtain relevant information while meeting them in person.

Table 3.2 Overview of the tests and procedures

Order	Tests of	Sub-test	Description
1	Working memory	NonWord	Phonological short-term memory
		SSpan	Executive working memory
2	Language aptitude	LLAMA B	Vocabulary learning
		LLAMA D	Sound recognition/sequence learning
		LLAMA E	Sound-symbol pairing
		LLAMA F	Grammatical inferencing
3	Implicit and explicit knowledge	EI	Implicit knowledge
		TGJT and WordM	Implicit knowledge
		UGJT	Explicit knowledge
		MKT	Explicit knowledge
		Cloze	Implicit knowledge
4	Background and learning experience	Questionnaire	Background and learning questionnaire of the NNS

Each of the participants was individually tested by the researcher. The testing time ranged between 2 to 2.5 hours for the NNS, and 1 to 1.5 hours for the NS. To reduce the possible influence of fatigue on performance, participants were allowed to take breaks between tests.

In the following sections, these tests and the questionnaire will be described in more detail. Test administration procedures and test scoring will also be presented.

3.4.2 Tests of implicit linguistic knowledge

3.4.2.1 Oral elicited imitation test

The EI test developed by Erlam (2006; 2009) was adapted in the present study. This test comprised 34 stimuli sentences, targeting the 17 linguistic structures listed above (see Section 3.3). The participants were asked to listen to some sentences, some of which were grammatical and some ungrammatical. Then, the participants were asked to indicate whether they thought the sentence that they had just listened to was true or not true. Finally, they were required to repeat the sentence. It was

hypothesised that the grammatical accuracy of the participants in sentence repetition would reflect their level of implicit knowledge. It was also hypothesised that if a participant had implicit knowledge of a target structure, he or she would make corrections in sentence repetition when the stimuli sentences were ungrammatical.

In order to maximise the possibility of accessing implicit knowledge, Erlam (2006; 2009) emphasised the following three aspects in test design and administration:

1) Focus on meaning

The primary focus of the participants should be on the meaning of the sentences. Empirically, the 17 target structures were made into 34 belief statements to require the participants to focus on meaning and comprehend what they had heard. These beliefs (i.e. stimuli sentences) were played to the participants and they then had to make a "belief choice" and indicate whether the opinion expressed in the sentence was true or not true for them. To make a belief choice, the participants would need to comprehend the sentences and process them for meaning.

In test instructions, the participants were told to repeat the sentence in correct English. That is to say, although the correction of grammatical errors in sentence repetition was regarded as evidence of the use of implicit knowledge, the participants were not asked to make error corrections and their attention was not explicitly drawn to the fact that some sentences were ungrammatical. In fact, the participants were never told that there would be both grammatical and ungrammatical sentences, and that they would need to correct ungrammatical sentences. However, this test included a training session in which the participants practised with some sentences. In this training session, participants were given both grammatical and ungrammatical sentences to practise with and each time were provided with model answers in correct English.

2) Reconstructive in nature

To reduce the likelihood of using rote memory in the completion of the EI, this test was designed to be reconstructive, that is, to minimise the possibility that

participants would repeat what they had heard verbatim. To this end, equal numbers of grammatical and ungrammatical stimuli sentences were included, and the inclusion of ungrammatical sentences was specifically used to test whether automatic correction of grammar errors would take place. In addition, to further enhance the reconstructive nature of this test, a delay between the presentation of the test sentences and the repetition was created by asking the participants to make belief judgements. If one failed to understand the sentence, it was less likely that he or she would be able to reproduce the sentence after a 3 seconds delay (McDade, Simpson, & Lamb, 1982). The delay between the presentation and the repetition of stimuli sentences in the EI enhanced the possibility that the participants were required to access their own L2 knowledge as they completed the test.

3) Time pressure

This test was implemented with time pressure so that it was more likely that implicit knowledge was accessed. Erlam's (2006, 2009) implementation of this time pressure was to present the stimuli sentences only once, and the participants were required to repeat each sentence without delay after their belief choice had been made.

The present study made some adjustments to Erlam's (2006, 2009) test. Firstly, instead of playing the stimuli sentences with a cassette player, this test was computerised using DMDX version 5.1.4.2 (Forster & Forster, 2003) and presented to the participants on an ASUS computer. The DMDX program also recorded the responses of the participants. The original recording of Erlam (2006, 2009) was transformed into digital files and then used as stimuli for the present study. Test instructions, practice sentences, and stimuli sentences all remained the same as in Erlam (2006, 2009). Participants sat with the researcher individually and were asked to follow the instructions on the computer screen (see Figure 3.3.1 for an example). However, instead of playing the test instructions aurally to the participants as in Erlam (2006, 2009), the present study presented the instructions visually. The test

procedure was therefore similar to that of Erlam (2006, 2009) but adapted. The adaptations will be explained in more detail in later paragraphs.

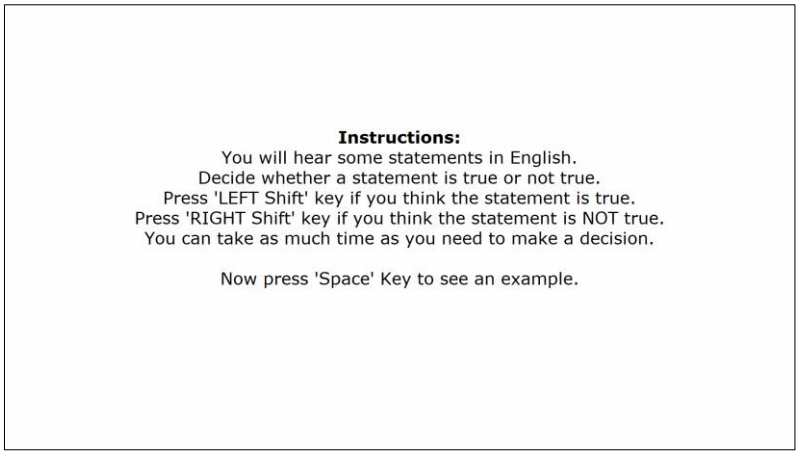

Figure 3.1 Screenshot of EI test instructions

The second change to the EI in the present study was to impose a 10 seconds time pressure for sentence repetition on top of the requirement of hearing each stimulus sentence only once and in real time. This decision was made because some scholars have argued that it is possible to access explicit knowledge even under time pressure (DeKeyser, 2003; Rebuschat, 2013). Therefore, to reduce the likelihood of the possible involvement of explicit knowledge, a tentative 10-second time limit was trialled in the pilot study (see Section 3.6) and then imposed.

The third change to Erlam's (2006, 2009) EI was to exclude the "Not sure" option in the belief choice questions. In Erlam (2006, 2009), belief choices were made with pen and paper. The participants were given an answer sheet and were asked to choose between "True", "Not true" and "Not sure" after they had listened to a test sentence. In comparison, in the computerised version of the EI in the present study, only two options, that is, "True" and "Not true" were provided. This decision was made because it was considered easier for the participants to press two designated keys (i.e. the left shift key for "True" and the right shift key for "Not true") rather than three different keys. Considering that the participants would need to press the space key as they proceeded with the EI, having more than two options

in the belief choice questions means that the participants would need to remember more keys for test completion, which could potentially cause difficulty for some participants. Moreover, it was considered that the use of two options (i.e. “True” and “Not true”) would not change the way that belief choices were made because the participants would still need to comprehend the sentence in order to make a belief choice. Therefore, since the use of two options in belief choice questions would probably be more user friendly for the participants, this adjustment was made.

In addition, the question “Did you notice anything special about the sentences?” was asked of every participant at the end of the EI. The main purpose of asking this question was to determine whether the participants were aware of the ungrammatical stimuli sentences or not. The answers of the participants to this question could provide some evidence of the degree of awareness involved in the repetition process. The question was purposefully phrased as an implicit open-ended question so that participants were not led to think of the grammaticality of the sentences. The researcher received a range of different answers, and answers that were relevant to grammar such as “Yes, the word order was wrong” were taken as evidence of being aware of grammatical errors. In contrast, answers that only addressed the content such as “There were a lot of sentences about the Royal Family.” were taken as evidence of the absence of awareness.

As in Erlam (2006, 2009), the EI test in the present study included a practice session so that participants were given a chance to familiarise themselves with the test procedure. In the practice, the participants went through the exact test procedure including receiving test instructions that were presented visually, listening to the stimulus sentence, making a belief judgement, and repeating the sentence. Model answers were provided in the practice, so that the participants would know what answers were expected. These model answers demonstrated how grammatical sentences were repeated and ungrammatical parts of the sentences corrected without any structural changes made to the original sentence. There were 8 sentences for the participants to practise with, and model answers were played to the participants

when the space key was pressed (Figure 3.2). In the practice session only, the participants saw the instruction sentence “You should have repeated the sentence in CORRECT English’ after each practice” (Figure 3.2). At the end of the practice, an instruction page was shown to remind the participants that they were required to make a belief judgement first, and then to repeat in “CORRECT English” (Figure 3.3). The word “correct” was capitalised in the practice session because the researcher wanted to reduce the likelihood that the participants repeat verbatim even when the stimuli sentences were ungrammatical, although this emphasis on “correct English” could have made the idea of grammatical correctness or incorrectness more salient in comparison to Erlam (2006, 2009). On the other hand, as the EI test had a primary focus on meaning and this test instruction was not shown during the main test, the participants were not constantly made aware that they would need to pay attention to grammatical correctness.

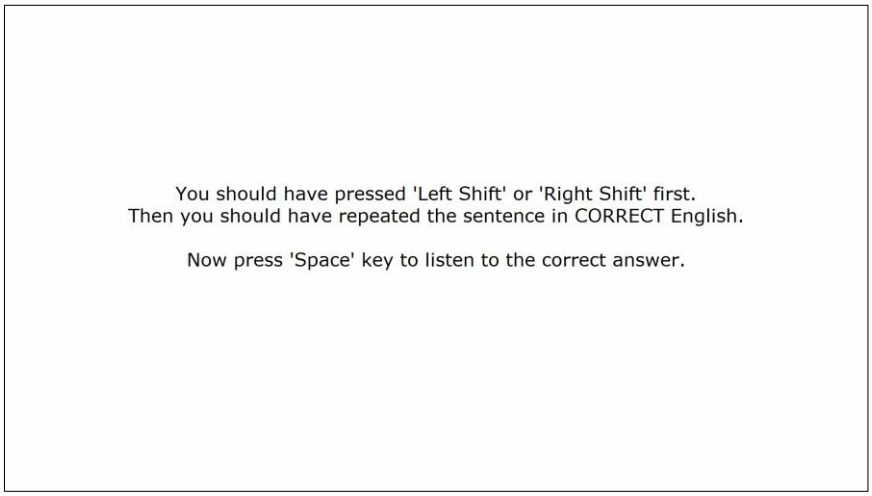

Figure 3.2 Screenshot of the EI instructions – practice

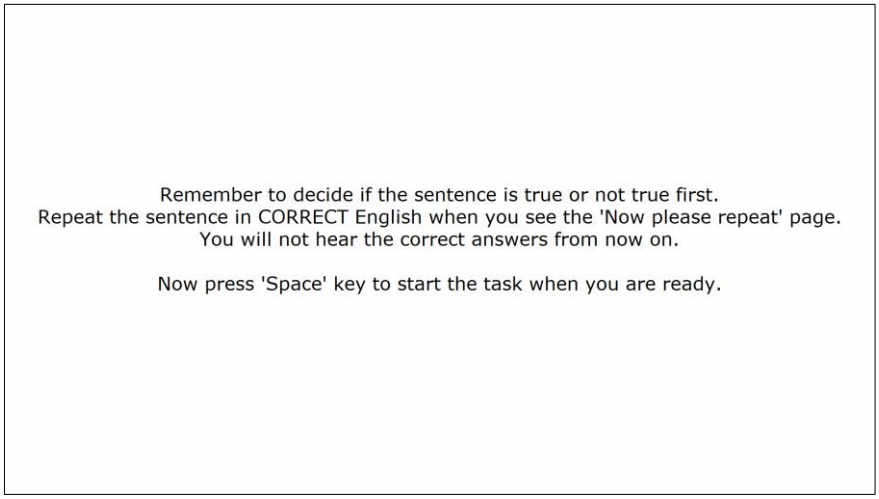

Figure 3.3 Screenshot of the EI instructions – end of practice

After the practice, the participants moved on to the test. The overall procedure of the EI in the present study was:

- The participants listened to a stimulus sentence. A "+" mark appeared at the centre of the screen as the sentence was playing (Figure 3.4).

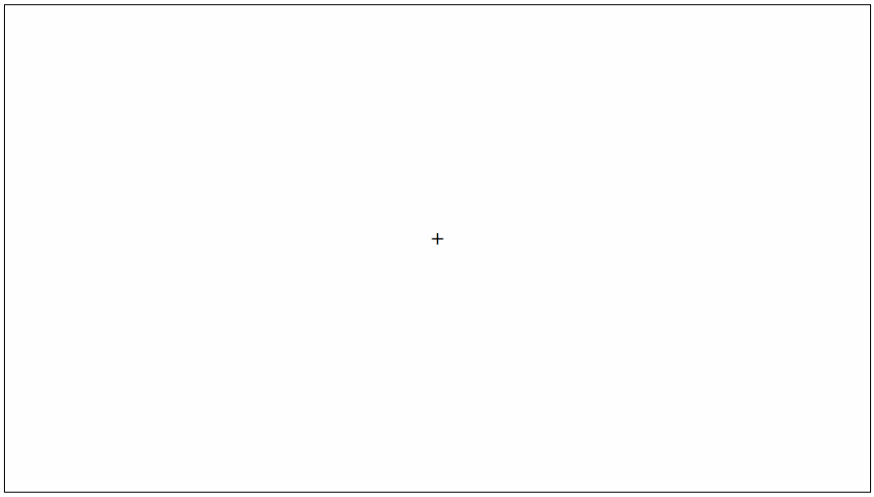

Figure 3.4 Screenshot of the EI – playing of stimuli

- The participants made their belief choice after they had listened to the sentence. They pressed the left shift key when they thought the sentence was true, or the right shift key when they thought the sentence was not true (Figure 3.5). There was no time limit for making a belief choice.

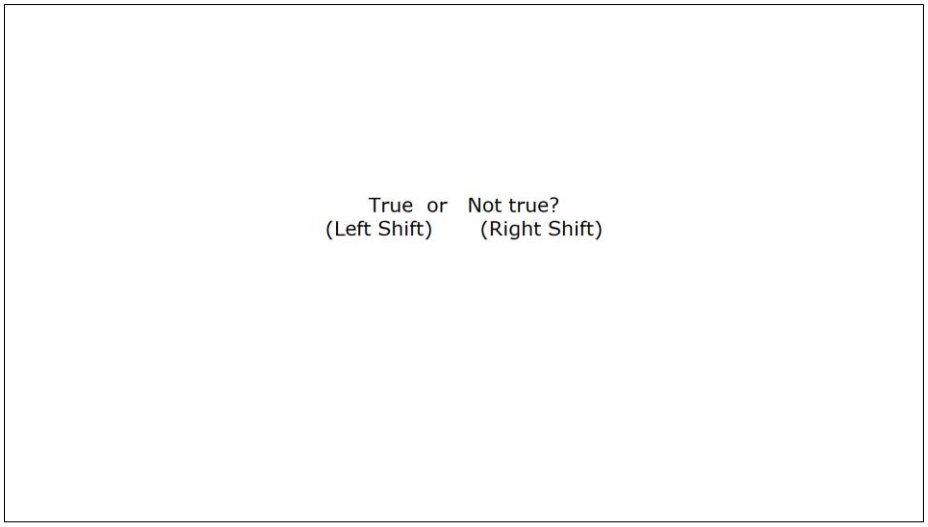

Figure 3.5 Screenshot of the EI – belief choice

- The participants were asked to repeat the sentence (Figure 3.6). The "repeat in CORRECT English" instruction no longer appeared in the main testing session of the EI. Instead, the participants saw "Now please repeat" on the screen, and were asked to indicate their completion by pressing the space key. A timer was started when a belief choice was made, and lasted for 10 seconds for each sentence repetition. After 10 seconds, the next sentence started to play.

Figure 3.6 Screenshot of the EI – sentence repetition

- After the repetition of all 34 sentences, the participants answered the awareness question asked by the researcher, that is, whether they noticed anything special about the test sentences.

Participants' sentence repetitions were transcribed by the researcher and then marked by two independent raters. Following Erlam (2006, 2009), scoring was based on the correctness of target structure (see Appendix 1, target structures are underlined and ungrammatical sentences are marked with an asterisk). In other words, scores were given as long as the target structures were produced correctly, and the grammaticality of the rest of the sentences did not influence scoring. Each correct repetition was awarded 1 mark, whereas incorrect repetition received no mark. Instances where the target structure was not produced didn't receive any mark. Regarding the answers to the awareness question of the participants, the researcher coded the data as either "aware" or "unaware" of the grammaticality of the test sentences.

3.4.2.2 Word monitoring and timed grammaticality judgement tests

In the present study, the WordM was embedded in the aural TGJT test. The decision that these two tests were integrated was based on the following two considerations. Firstly, sentences in these two tests needed to be presented aurally, and this shared mode of presentation (i.e. aural stimuli) made it possible for the researcher to combine the two. Secondly, grammatical and ungrammatical sentences were included in both tests, and the same set of sentences could be used. Therefore, the test sentences of the TGJT developed by Loewen (2009) were adapted. Thirdly, the combination of these two tests would shorten the overall test-taking time for each participant, so that the possible interference of fatigue in test completion could be reduced. Therefore, it was decided that the WordM would be integrated into the TGJT.

A number of studies have used the WordM as a measure of implicit knowledge

(Granena, 2013b; Suzuki & DeKeyser, 2015; Suzuki, 2017; Vafaee et al., 2017). These studies provided some evidence that the WordM could be a good measure of implicit knowledge because it reflects automatic use of linguistic knowledge where explicit knowledge and strategy use is minimal (Kilborn & Moss, 1996). In light of this, the WordM in the present study was programmed using DMDX 5.1.4.2 and delivered to the participants using an ASUS computer. In this test, a word (i.e. the target word in each sentence) was presented visually to the participants and in the meantime a sentence that contained this word was played aurally. The participants were required to press the space key as soon as they heard the target word (Figure 3.7). A timer was triggered at the onset of the delivery of the target word, and the DMDX program automatically saved the response time (RT) in milliseconds, that is, the time taken from the onset of the target word to the time that the space key was pressed. Both grammatical and ungrammatical sentences were included in this test, and a so-called grammaticality sensitivity index (GSI) was calculated by subtracting the RTs of the grammatical sentences from the ones of the ungrammatical sentences.

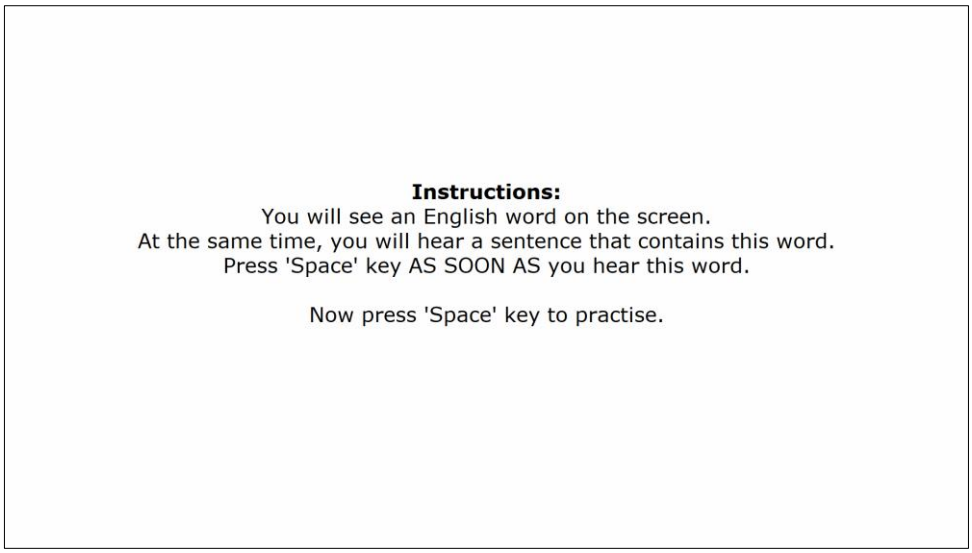

Figure 3.7 Screenshot of the WordM – Instructions

The inclusion of both grammatical and ungrammatical sentences was central to the underlying rationale of using this test as a measure of implicit knowledge,

because one's sensitivity to grammatical errors was reflected by the difference in RTs to grammatical and ungrammatical sentences. If implicit knowledge was being used in the monitoring process, there should have been a delay in RTs when listening to ungrammatical sentences (Granena & Long, 2013a; Marslen-Wilson & Tyler, 1980). That is to say, the lengthening of RT in ungrammatical sentences demonstrates the noticing of grammatical errors, which in turn shows that implicit knowledge is being activated.

Therefore, in order to examine participants' sensitivity to grammatical errors, the target words to be monitored always appeared after the target structures of the sentences (Jiang, 2013; Peelle, Cooke, Moore, Vesely, & Grossman, 2007). For example, the same target structure of the two sentences below was the use of third-person singular. The target words in the WordM test were "Auckland" in sentence 1 and "fish" in sentence 29, which appeared after the target structure "work" and "eats" respectively. If someone was sensitive to the misuse of third-person singular, he or she should have taken longer to monitor the target word in the ungrammatical sentence 1 than the grammatical sentence 29.

1. * Kim lives in Hamilton but work in **Auckland** this year.

29. Keiko eats a lot of **fish** every week.

The choice of the target words was also determined by their acoustic salience because words that were not acoustically salient, such as the function word "in" in the above example sentences, would be very difficult to monitor. Therefore, the target words in the WordM were mostly content words that were acoustically salient (Jiang, 2013). The only exceptions were sentences 31 and 42 (see below). In these two sentences, although the target words were not content words, they were acoustically salient. The target words in this test were highlighted in bold type in the sentences (see Appendix 2).

31. Bill wanted to know where I **had** been.

42. * The population of New Zealand was increased **between** 1990 and 2000.

Some changes to the test sentences in Loewen's (2009) study were made in the present study. Firstly, typical English names such as Kim and David were used to replace Keum and Liao in Loewen (2009). This modification was based solely on the personal speculation of the researcher because she thought that this would probably make it easier for the participants to judge the grammaticality of the sentences. Secondly, some changes to sentence structures were made considering the needs of the WordM. In order to allow some time for word monitoring, the target words in the WordM could not be the last word in each sentence. Therefore, without making changes to the target structures of the sentences developed by Loewen (2009), one of the following two kinds of modifications was made when necessary. The first kind of modification was to make some sentences longer by adding adverbials. For example, sentence (a) below was the original sentence in Loewen (2009), and sentence 5 was the sentence used in the present study. The target structure of this sentence was the use of possessive -'s, which occurred at the end of the original sentence. In order to allow some time for monitoring the target word "teacher", an adverbial phrase "two days ago" was added. For the same purpose of allowing for some word monitoring time, the second kind of modification was to alter the word order of the original sentence in Loewen (2009). For example, in sentence (b), the target structure was the use of ergative verbs. This sentence was changed to sentence 42 in the present study so that the target word "between" would occur after the target structure.

a. * Keum went to the school to speak to her children teacher.

5. * Kim went to the school to speak to her children **teacher** two days ago.

b. * Between 1990 and 2000 the population of New Zealand was increased.

42. * The population of New Zealand was increased **between** 1990 and 2000.

However, the above listed modifications could not be applied to the sentences that target the use of question tags. For example, in the following sentence 62, the target structure occurred at the very end of each sentence, and it was impossible to add an adverbial. Therefore, these sentences were regarded as filler sentences and the target words were chosen randomly. The RTs of these sentences were excluded in data analysis.

62. That book isn't very interesting, is it?

The stimuli sentences of the WordM and TGJT were read and recorded by the researcher (a female non-native speaker of English) who is proficient and does not have a strong accent. However, it was possible that the way that the researcher read the sentences was easier for the NNS participants to follow because both of them had the experience of learning English in China. The audio files of each sentence were split into two parts at the onset of the target words. The splitting of the audio files was a necessary step because a timer needed to be placed at the onset of the target words. However, this splitting of the audio files was not noticeable to the participants.

As they monitored the target words in the WordM, the participants completed the TGJT. In other words, as the participants listened to the sentences that contained the target words for monitoring, they were also listening and judging the grammaticality of the sentences of the TGJT. Again, as a measure of implicit knowledge, the TGJT was time pressured. The participants were given 5 seconds to decide if the sentence they had just heard was grammatical or ungrammatical. After 5 seconds, the next target word would appear on the screen and the next sentence would start playing. For the TGJT, a timer was triggered from the offset of the aurally presented sentence and was stopped when a decision was made about the grammaticality of the sentence (i.e. the designated key was pressed, either the left shift key for grammatical sentences or the right shift key for ungrammatical

sentences). The accuracy of the judgement of the participants was recorded by the DMDX program.

It was decided that the stimuli sentences of the TGJT were presented aurally to the participants because recent research findings have indicated that aural TGJTs are more likely to access implicit knowledge than written TGJTs (Kim & Nam, 2017; Spada, Shiu, & Tomita, 2015). Through comparison of the scores of an aural TGJT and a written TGJT, both Kim and Nam (2017) and Spada et al. (2015) arrived at the conclusion that the use of aural stimuli taps into "more advanced, stronger implicit knowledge" (Kim & Nam, 2017, p. 24). This is probably because when completing an aurally presented test, there is a higher requirement for simultaneous language processing ability compared to the written mode where the visual input is available for reprocessing (Danks, 1980).

The use of an aural test had another advantage, that is, to decrease the possible interference of reading speed in a written TGJT. In Loewen (2009), the allocated time for the NNS in the TGJT was calculated by adding 20% to the mean judgement time of 20 NS. This also implicitly imposed a requirement for reading speed that is fast enough to cope with the allocated time. The NNS participants who had slow reading speed might be faced with the risk of not finishing the entire sentence, which in turn would result in relatively low scores for this test. Arguably, a slow reading speed can be taken as a reflection of the lack of sufficient implicit knowledge, and low accuracy in judgement is a result of this lack of implicit knowledge. However, alternatively, it is possible that variations in reading speed influence judgement accuracy as an independent variable. This possible interference can be avoided if stimuli sentences are presented aurally, by which a relatively equal chance is given to the participants to process the entire sentence while also maintaining a high requirement for automatic language processing. Therefore, it was decided that the aural mode would be used in the TGJT and a judgement time of 5 seconds was allocated to each sentence.

The 5 second time pressure was a somewhat arbitrary choice. Considering that

the participants were required to complete two tasks consecutively and that they would all hear the entire sentence before being asked to make a grammaticality judgement, the researcher decided to impose a unitary time constraint of 5 seconds. This time pressure was trialled in the pilot study and found to be adequate (see Section 3.6).

Again, the participants were given opportunities to practise and to familiarise themselves with the test procedure. Ten sentences were given as practice sentences, and then the participants moved on to the test. The overall procedure of the WordM and TGJT was:

- The participants saw a target word on screen and listened to a sentence that contained the target word at the same time (Figure 3.8). The timer was triggered at the onset of the playing of the target word to measure the RT of word monitoring.

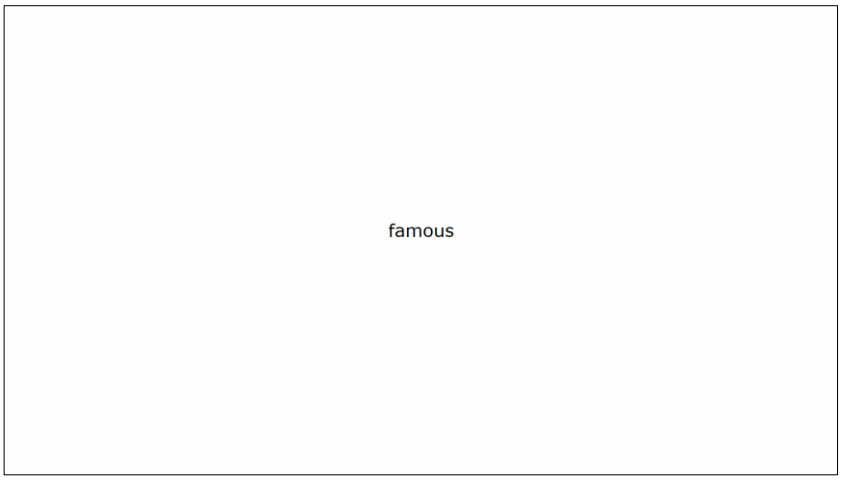

Figure 3.8 Screenshot of the WordM – Target words

- The participants pressed the space key to indicate that they heard the target word. The timer was stopped when the space key was pressed.
- The participants were asked to judge the grammaticality of the sentence (Figure 3.9). The timer was triggered again at the end of the sentence for

the TGJT.

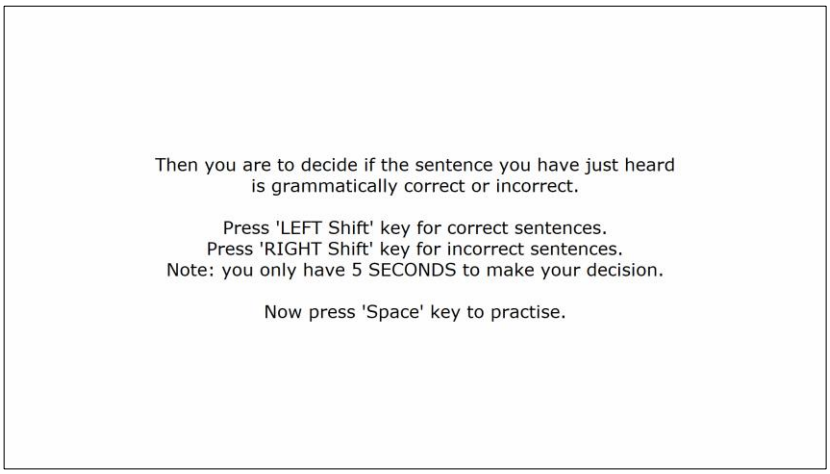

Figure 3.9 Screenshot of the TGJT – Instructions

- The participants made a grammaticality judgement by pressing the left shift key for grammatical sentences and the right shift key for ungrammatical sentences (Figure 3.10). An allocated time of 5 seconds was given to the participants to make their judgement for each sentence.

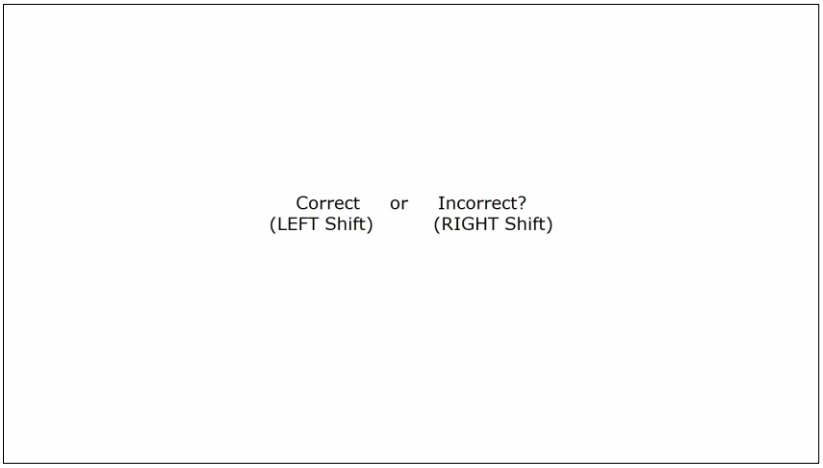

Figure 3.10 Screenshot of the TGJT – Judgement

- The next target word appeared on screen after 5 seconds

The RT data of the WordM was automatically saved in the DMDX output file in the form of milliseconds. Instances where participants failed to respond in the WordM were coded as missing data. The scoring of the TGJT was dichotomous: correct judgements received 1 point, while incorrect or no responses received 0 points.

3.4.2.3 Cloze test

The cloze test was administered as a measure of implicit knowledge. Specifically, this test aimed to examine English collocational knowledge of the NNS. The term collocation refers to "an aspect of lexical cohesion which embraces a 'relationship' between lexical items that regularly co-occur" (Halliday & Hasan, 1976, p. 288). For example, it is common in English to say "make the bed" and "do homework", rather than "tidy the bed" or "write homework". This type of knowledge is implicit in nature, which comprises a significant part of native speaker competence (Hill, 2000), and also second language competence (Wray, 2000). In other words, the more collocational knowledge one possesses, the more fluent and accurate one's L2 use will be; and the closer this person's language use resembles that of NS. Therefore, the present study adopted the cloze test of Davies (1991) as a measure of collocational knowledge. It was hypothesised that the performance of the NNS in this test would reflect how far their use of the L2 deviated from NS norms.

This test comprised 20 independent sentences, within which 44 gaps were included (see Appendix 3). The test was administered without time pressure with pen and paper. Participants were instructed to fill in the gaps to make meaningful and grammatical sentences. They were told that there were many possible answers and only one word was allowed for each gap.

In his book, Davies (1991) provided some answers to this test (see the first column in Appendix 4). However, as there were multiple ways to complete each sentence, it was decided that some other possible answers would be considered in the

present study. Therefore, three university linguistics lecturers (NS of English) were asked to give possible answers. These answers, together with the ones provided by Davies (1991) made up a list of acceptable answers (see the second column in Appendix 4). The participants received 1 point for each correct answer that was in this list.

There were cases when the answers of the participants were outside the list. The researcher gathered these answers and asked one of the three university lectures to mark these answers according to their appropriateness and collocational strengths. The rater mainly considered the strength of collocation within each sentence, and thus answers that had weak collocations (unusual) within the sentences were not accepted.

Some sentences had more than one gap. In the marking of these sentences, the following rule was followed. If the gaps were in the same clause, the participants would need to provide correct answers for every gap in order to get the points. Otherwise the participants would receive 0 points. For example, in sentence 4 below, one accepted answer was not/but. The model answer itself showed that in the context of this sentence, “not” and “but” would need to be used in conjunction with each other. Therefore, answers that were partly correct were not accepted. That is to say, in such sentences, the participants either received 2 points or no point at all. However, if the gaps were in separate clauses, 1 point was given for each correct answer. For example, in sentence 11, the participants received 1 point if “not” was put in the first gap. Even if the second gap was not filled in correctly, the participants were still rewarded for individual correct answers.

4. He would have to ask his wife’s permission before lending the case because it was_____ his _____ hers.

11. As I am a little deaf, I find that I can _____ always hear what is said _____ the phone.

3.4.3 Tests of explicit linguistic knowledge

3.4.3.1 Untimed written grammaticality judgement test

Unlike the tests of implicit knowledge that imposed a time pressure, the tests of explicit knowledge allowed unlimited time in test completion. In the UGJT, the absence of time pressure would most likely provide an opportunity for the NNS to resort to their explicit knowledge (R. Ellis, 2005; R. Ellis & Loewen, 2007; R. Ellis et al., 2009; Han & Ellis, 1998; Loewen, 2009). However, it should be noted that participants did not necessarily need to use explicit knowledge to complete this test. The NS participants, whose knowledge was largely implicit, and some near-native NNS with a relatively high level of implicit knowledge, might still use implicit knowledge to make a grammaticality judgement.

Again, the UGJT was programmed using DMDX 5.1.4.2 and presented to the participants on an ASUS computer. Following Loewen (2009), the UGJT included a measure of certainty, that is, how certain the participants were about their grammaticality judgement for each sentence. However, unlike Loewen (2009) who asked the participants to type in their certainty in a percentage (i.e. __ %), the present study asked the participants to indicate their certainty on a scale of 1 to 5 (i.e. pressing one of the number keys of 1 to 5), with 1 being not certain at all and 5 totally certain. This change was made because the DMDX program did not allow for the typing in of information, so that the level of certainty was put on a Likert scale format for the participants to choose from (Figure 3.11).

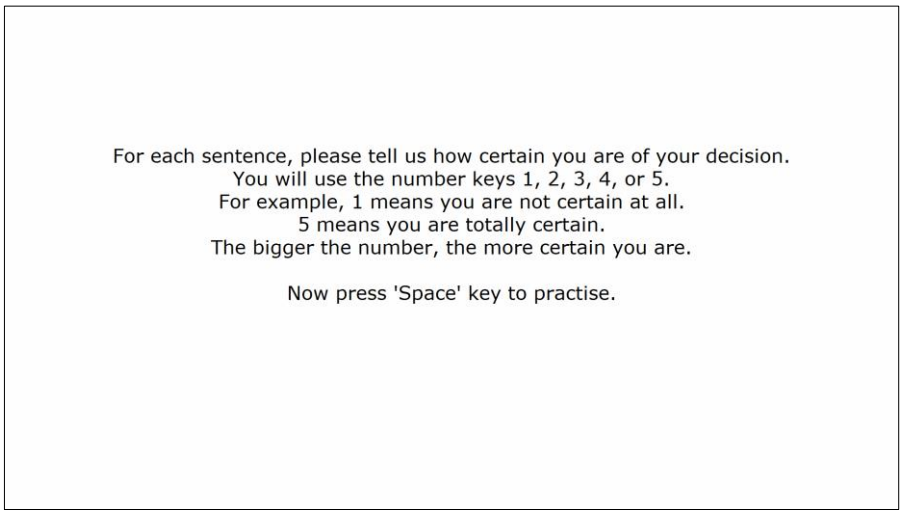

Figure 3.11 Screenshot of the UGJT – Instructions

Considering the overall length of the entire test battery (approximately 2 to 2.5 hours for the NNS), the present study asked the participants about their source of judgement (i.e. by "rule" or by "feel") at the end of the UGJT rather than asking this question for each sentence. After the participants completed the grammaticality judgement of the 68 sentences and indicated their level of certainty, the researcher asked them whether they made their judgement by intuition, by following grammar rules, or by both following intuition and grammar rules.

The UGJT was completed after the TGJT. Test sentences in the UGJT were identical to those used in the TGJT but were presented to the participants in written form. By using identical sentences, the participants were compelled to think about the grammaticality of these sentences a second time. This would probably increase the likelihood of resorting to more explicit knowledge, although we cannot be completely confident that explicit knowledge was the pure source of their judgement.

A practice session was included in the UGJT as well. Ten sentences were given to the participants to practise with so that they could get familiar with the test procedure. The answers of the participants were marked dichotomously as correct or incorrect. A correct judgement received 1 point, and an incorrect judgement received 0 points.

The procedure of the UGJT was:

- The participants saw a sentence on the screen and were asked to decide if the sentence was grammatical or ungrammatical by pressing the left shift key for grammatical sentences and the right shift key for ungrammatical sentences (Figure 3.12).

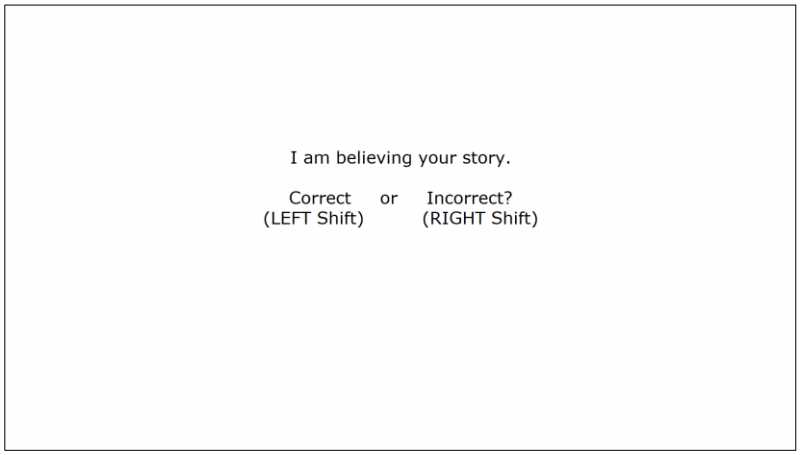

Figure 3.12 Screenshot of the UGJT – Judgement

- The participants were asked to indicate their level of certainty by pressing one of the number keys of 1 to 5. Larger numbers indicated higher levels of certainty, with 1 being the least certain, and 5 being completely certain (Figure 3.13).

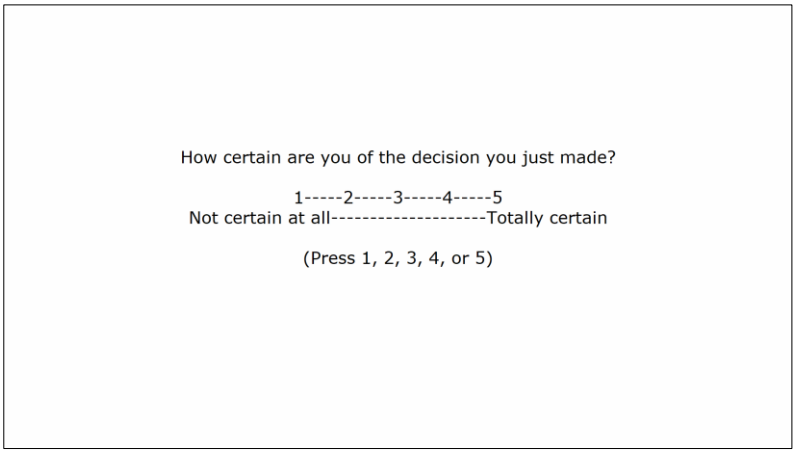

Figure 3.13 Screenshot of the UGJT – Certainty

- At the end of the UGJT, the researcher asked the participants about their source of judgement, that is, whether they made their judgement by intuition, by grammar rules, or by a combination of intuition and grammar rules. Each participant gave an oral response to this question.

3.4.3.2 Metalinguistic knowledge test

Following the UGJT, the participants were asked to complete a MKT. This test was adopted from Elder (2009) and designed as a measure of metalinguistic knowledge. Metalinguistic knowledge is a type of knowledge that involves "explicit declarative facts (whether rules or fragments of information) that a person knows about language" (Elder, 2009, p. 114). This type of knowledge is analytic in nature and is often operationalised as learners' ability to describe and explain linguistic rules (Elder, Warren, Hajek, Manwaring, & Davies, 1999; Roehr, 2006; 2008). That is to say, NS do not necessarily possess this knowledge, while second language learners may have it to some extent, especially if they have received explicit instruction in an L2 classroom.

Unlike the other test of explicit knowledge, this test did not require participants to judge grammaticality. Rather, participants were required to choose the correct grammatical explanation of errors, to match grammatical terms from a given passage, or to label parts of speech in sentences (Alderson et al., 1997). During this process, the participants would need to understand metalinguistic terminologies so that the test could be completed, yet the participants were not required to produce accurate metalanguage.

Considering that the NNS in the present study may not be familiar with English metalanguage because they probably had learnt grammar in an EFL context using Mandarin Chinese, an annotated Mandarin version of the MKT was prepared for the NNS (see Appendix 5). The original English version remained the same and was used for the NS.

In the annotated MKT test, all the instructions were written in simplified Chinese and the metalanguage in the questions was translated into corresponding Mandarin Chinese metalanguage. The translation was put in brackets next to the original text. For example, in question 1 of the MKT, the metalanguage "modal verbs", "preposition", "base form" and "infinitive" were translated into the corresponding metalanguage in simplified Chinese.

1. You must to wash your hands before eating.
   a. Modal verbs (情态动词) should never be followed by a preposition (介词).
   b. After 'must' use the base form (原形) of the verb not the infinitive (动词不定式).
   c. Modal verbs (情态动词) should never be followed by a preposition (介词).
   d. After 'must' use the base form (原形) of the verb not the infinitive (动词不定式).

The MKT consisted of three parts. In part 1, the participants were asked to answer multiple choice questions and choose a correct explanation for the grammatical errors in question. A total number of 17 questions corresponding to the 17 target structures (see Section 3.3) were included. In part 2, a short passage was given to the participants, and the participants were asked to find examples from the passage that corresponded to given grammatical terms. In part 3, the participants were asked to underline the requested sentence elements.

Participants were given unlimited time to complete this task with pen and paper. Some NS found it really difficult since they had never been exposed to linguistic terminologies before. In this case, the researcher encouraged the participants to guess an answer and do as much as they could for parts 2 and 3. The NNS were more familiar with the test format, and they were also encouraged to make a guess if they were unsure about their answers.

The scoring of the MKT was based on the number of correct answers, with 1 point given to each correct answer and 0 points for incorrect answers. In part 3, most

of the participants were unable to provide accurate answers. For example, in the question below where the participants were asked to underline the subject of this sentence, most participants underlined “Joe” instead of “Poor little Joe”. In this case, 1 point was still given because Joe was the head word of the subject phrase. Although the complete subject phrase was not identified, the underlining of the head word “Joe” showed that the participants had some syntactic knowledge.

1. Poor little Joe stood out in the snow. (Subject)

3.4.4 Tests of working memory

The present study included two tests of working memory because it has been suggested that working memory constitutes a part of language aptitude (Miyake & Friedman, 1998; Robinson, 2002a; 2005; Sawyer & Ranta, 2001; Wen & Skehan, 2011). Two types of working memory ability, that is, phonological short-term memory and executive working memory, have been investigated by previous research due to their relevance to second language acquisition (Skehan, 2015b; Wen, 2015; 2016). In the present study, a NonWord test was used as a measure of phonological short-term memory, and a listening SSpan test as a measure of executive working memory.

One key characteristic of working memory is that it is not language independent (Kane et al., 2004; Osaka & Osaka, 1992; Osaka, Osaka, & Groner, 1993; Trofimovich, Ammar, & Gatbonton, 2007). That is to say, the test of working memory capacity of L2 learners should be administered in their L1 to avoid any interference of non-automatized L2 knowledge, especially when the proficiency level of the L2 is low (Mackey, Adams, Stafford, & Winke, 2010). Therefore, the two working memory tests were administered in Mandarin to the NNS. The NS did not take these tests.

3.4.4.1 Non-word repetition test (NonWord)

In this test, the participants were asked to listen to some non-words and then to recall them. This test has been commonly used as a measure of phonological short-term memory, which is particularly relevant to the phonological loop in working memory. The phonological loop is a sub-system under the central executive, which is responsible for the storage and rehearsal of phonological information. The rehearsal process happens in real time and is similar to subvocal speech, which in turn restricts the storage capacity (Kormos & Sáfár, 2008). This means that after hearing several items, the very first few items will no longer be available for rehearsal. Therefore, tests of the phonological loop are also often referred to as tests of short-term memory.

The present research used Zhao's (2015) Mandarin non-word repetition test. There were 24 trials in this test, which were made up of 48 one-syllable Mandarin non-words (see Appendix 6). A trial here is similar to a question in a test in the sense that the participants complete this test trial by trial. All of the non-words were pronounceable but did not have characters representing them. In other words, the non-words were all factitious, meaningless and irrelevant to each other. The number of words in each trial started from 2 and increased by 1 every three trials.

The stimuli non-words of this test were read and recorded by the researcher, who is a female native speaker of Mandarin. Noise was reduced by using editing software so that the voice quality was ensured. All the non-words were played at the fixed pace of 1 second per word.

The participants took this test individually, and their responses were audio-recorded using an ASUS computer. The researcher played the 24 trials one by one to the participants and guided them throughout the test. Each trial was played only once, and the participants were asked to repeat the non-words that they had just heard immediately after the trial of words had been played. A practice session was implemented, in which the participants were given opportunities to practise on six trials (three two-word trials and three three-word trials). After the practice, the

non-word span test began with two-word trials and all participants were required to complete all 24 trials. Each participant was encouraged to remember as many non-words as possible but was not required to remember the non-words in the sequence that these non-words were played. The researcher emphasised that in order to get 1 point, the participants would need to correctly pronounce the initial, the final and the tone.

Two native speakers of Mandarin marked the responses of the participants by listening to their recordings. The raters met in person and were given transcripts of the test stimuli. They then listened to participants' recordings and ticked the words that were correctly produced. Transcriptions of participants were not made in advance so that any possible influence of inaccurate transcription would be avoided[1]. The two raters agreed with each other most of the time, but there were times when an agreement couldn't be reached. In that case, they either gave or did not give a point based on their own auditory judgement.

The scoring of this test followed the percentage method because research has shown that such proportion scores have an advantage over other scoring methods of working memory tests, such as counting the absolute number of words correctly repeated (St Clair-Thompson & Sykes, 2010). Each correct repetition of a non-word received 1 point, including the correct repetition of the tone, initial and final. Incorrect repetitions of tone, initial, or final were not given any point. Then a total score was given to each trial based on the number of non-words correctly repeated by the participants. This total score was then divided by the total number of words in each trial to form a percentage score for each trial. Finally, these percentage scores were averaged out across trials to form the final score of this test.

[1] Even for some native Mandarin speakers, it is very difficult to differentiate certain sounds due to dialectal influence, such as the alveolar unaspirated affricate z [ts] and retroflex unaspirated affricate zh [t͡ʂ]; the syllable finals -n [n] and -ng [ŋ].

3.4.4.2 Listening sentence span test

A Mandarin listening SSpan test was used as a measure of executive working memory. Unlike the NonWord test that only assessed storage capacity, this complex span test examined both the processing and storage capacity of working memory. In this test, the participants were asked to do two things simultaneously: 1) to listen to sets of sentences and decide the plausibility of each sentence (whether it made sense or not), and 2) to remember the last word of each sentence as they listened. They then had unlimited time to write down or say the words that they remembered after listening to one set of sentences. The number of sentences in each set varied, and sets of sentences were normally randomized.

The test was adopted from Zhao (2015) and was a modified version of the test described in Lewandowsky, Oberauer, Yang, and Ecker (2010). The original test was developed as part of a larger working memory test battery (Lewandowsky et al., 2010), which was validated with participants in Taiwan and Australia. Considering differences in the norms of expression of Mandarin in Taiwan and mainland China, Zhao (2015) revised 96 sentences and piloted them with 10 native Mandarin speakers in mainland China. Finally, a total of 72 sentences were selected and were randomly arranged into 16 sets. The number of sentences in each set ranged from three to six. These sentences are shown in Appendix 7.

Test materials were read and recorded by the researcher. Recording quality was ensured by using software to reduce noise. All the sentences were programmed with DMDX version 5.1.4.2 and presented to the participants aurally on an ASUS computer.

This test was administered after the NonWord test. The researcher briefly introduced how the test should be done, and then guided the participants to follow the instructions on the computer screen. She also made sure that the participants fully understood the procedure and only started writing down or saying the words after listening to the entire set of sentences. In cases where participants failed to write down all the words, the administrator encouraged them to move on to the next

set.

The majority of the participants completed this test by writing down their answers on the answer sheet. However, a few participants who immigrated to New Zealand before the age of 10 were unable to write in Mandarin Chinese. So these participants said the words that they remembered and the researcher wrote them down. The following figure gives an example of the test instructions (Figure 3.14).

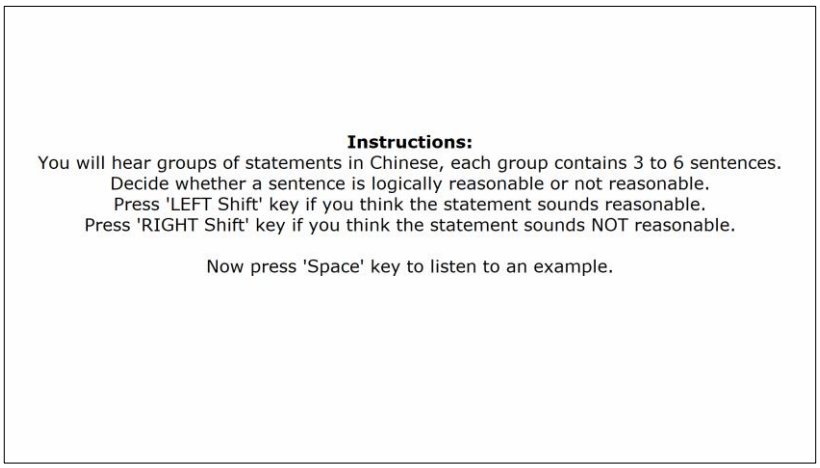

Figure 3.14 Screenshot of the SSpan – Instructions

Again, a practice session was implemented to familiarise the participants with the test procedure. In the practice session, the participants practiced with three sets of sentences, one with two sentences and the other two with three sentences. Then the participants moved on to the main test, and the procedure was:

- The participants listened to a set of sentences. While a sentence was playing, a "+" mark appeared at the centre of the screen to focus the attention of the participants (Figure 3.15).

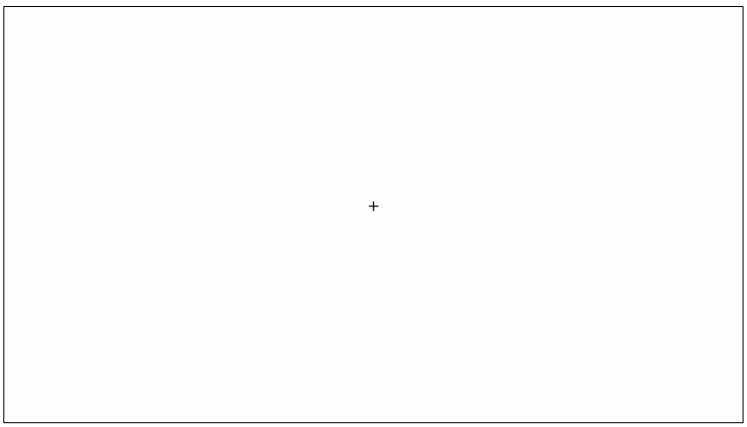

Figure 3.15 Screenshot of the SSpan – Playing of sentences

- The participants decided the plausibility of each sentence by pressing the left shift key if they thought the sentence was reasonable or the right shift key if the sentence was not reasonable (Figure 3.16). Once this decision was made, the next sentence in the set started to play and the participants saw the "+" on screen again.

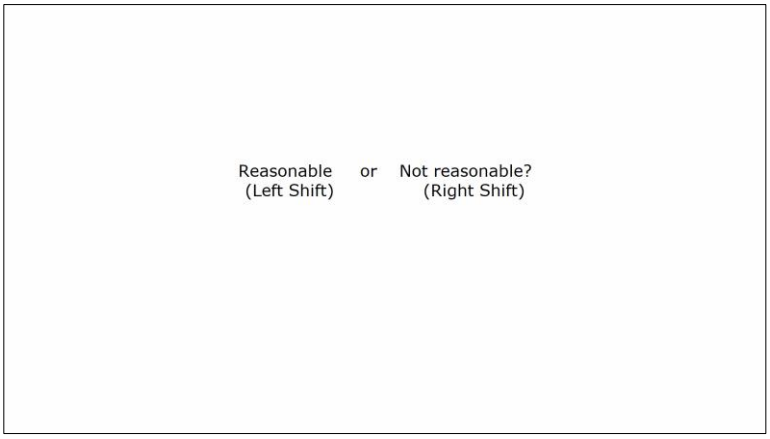

Figure 3.16 Screenshot of the SSpan – Plausibility judgements

- The participants attempted to remember the last words of the sentences they listened to as they made plausibility judgements. They were given instructions at the beginning of this test (Figure 3.17) and were told to remember as many words as possible.

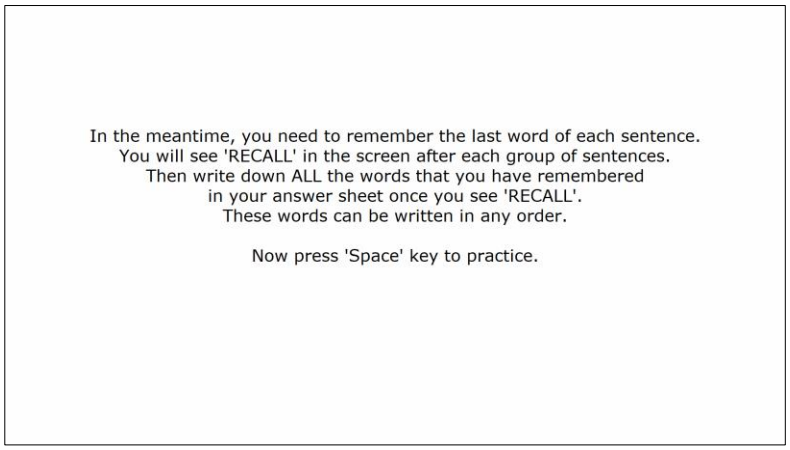

Figure 3.17 Screenshot of the SSpan – Recall instructions

- The participants wrote down/said the words that they remembered after they listened to a set of sentences (Figure 3.18).

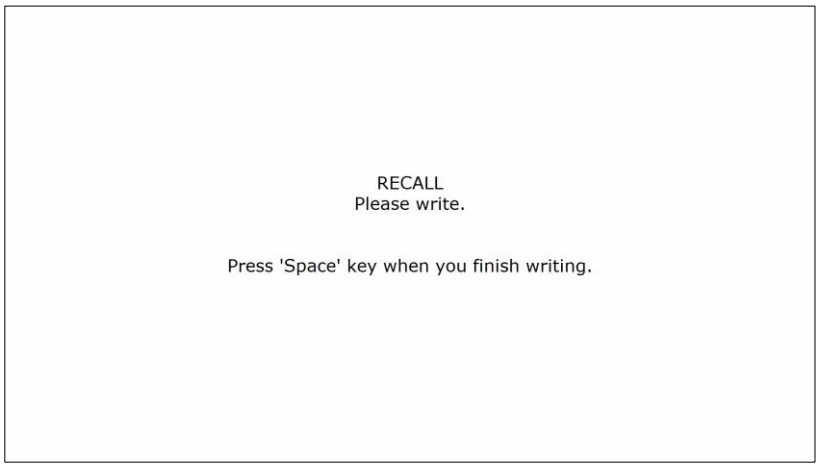

Figure 3.18 Screenshot of the SSpan – Recall

The scoring of this test followed the same percentage method as the NonWord test. For each word that was correctly recalled, 1 point was given. A total mark was given for the number of words correctly recalled for each set of sentences. Then this total mark was divided by the total number of words in each set to form a percentage score for each set. At last, these percentage scores were averaged out across all sets of sentences to form the final score.

3.4.5 Tests of language aptitude: the LLAMA test battery

The LLAMA language aptitude test (Meara, 2005) was administered as a measure of participants' language aptitude, and only the NNS participants completed this test. In recent years, this test has been used by a growing number of researchers in the field (e.g. Abrahamsson & Hyltenstam, 2008; Granena & Long, 2013b; Yilmaz, 2013). The primary strength of this test is the use of picture stimuli, which makes it a language independent measure of cognitive abilities. This is highly desirable for an aptitude test, therefore any confusion between language proficiency and language learning abilities can be avoided. It is also available freely in a readily built format that can be administered easily.

In order to make the test interesting and to maintain participants' enthusiasm, the researcher told the participants that they were going to learn a new language through the four sub-tests, namely LLAMA B, D, E and F. The LLAMA programme automatically produced scores in the test panel once a test was completed, so the participants could see these scores. The researcher also showed the participants the grade scores (i.e. below average, average, good, excellent) of each sub-test according to Meara's (2005) classification. In cases where some participants received relatively low scores in certain sub-tests, the researcher made efforts to reassure these participants and pointed out that he or she still had the possibility of learning English well. By doing so, the researcher tried to engage the participants in this test and minimise the potential negative influence of being discouraged by low scores in some sub-tests.

Before the test was administered, the researcher explained the test requirements to the participants in Mandarin and made them aware that there were no practice sessions in the four sub-tests. Instead, there was a learning phase and a testing phase in each sub-test. However, participants were required to complete different tasks in each sub-test, which are explained in the following paragraphs.

LLAMA B was a test of vocabulary learning. In this test, the participants were asked to learn 20 new words in a 2-minute study phase, and they learnt these words

by clicking on the pictures presented to them (Figure 3.19). The participants were asked to learn as many words as possible and were told that they could click on each picture as many times as they wanted. There was a timer in the centre of the LLAMA B panel, showing the remaining time of the study phase.

Figure 3.19 Screenshot of LLAMA B

After the study phase, the participants entered the testing phase. In the testing phase, words appeared at the centre of the panel and the participants were asked to match the words to the picture stimuli by clicking on the corresponding picture. The LLAMA B program automatically gave feedback for the answers of the participants, with a ding sound indicating correct answers and a bleep sound incorrect answers. After the completion of the testing phase, the program automatically produced a score of this sub-test.

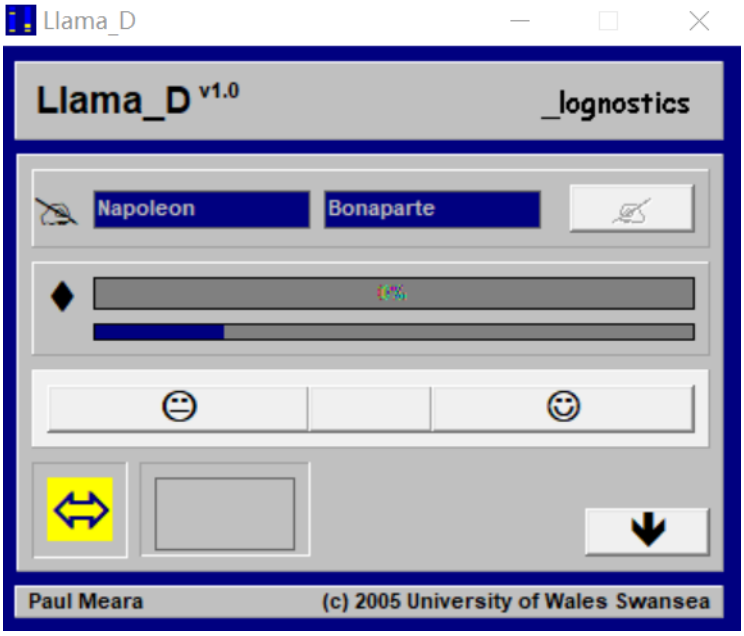

Figure 3.20 Screenshot of LLAMA D

LLAMA D (Figure 3.20) was a test of sound recognition. Again, this sub-test included a study phase and a testing phase. In the study phase, the participants were asked to listen to a string of 10 words of an artificial language. Immediately after the study phase, the participants moved on to the testing phase. In the testing phase, more artificial words were played to the participants. The words played in the testing phase were comprised of words that were played in the testing phase and some other artificial words that the participants had never heard before. The participants were asked to decide whether the words in the testing phase were words that they had heard before or words that were new to them. They indicated their choice by clicking on the button with a smiley face (i.e. the words had been heard in the study phase) or the button with a sad face (i.e. new words that had not been heard before). Again, feedback for the answers of the participants was given in the form of a ding sound (correct answers) or a bleep sound (incorrect answers) in the testing phase. At the end of the testing phase, the LLAMA D program automatically produced a test score.

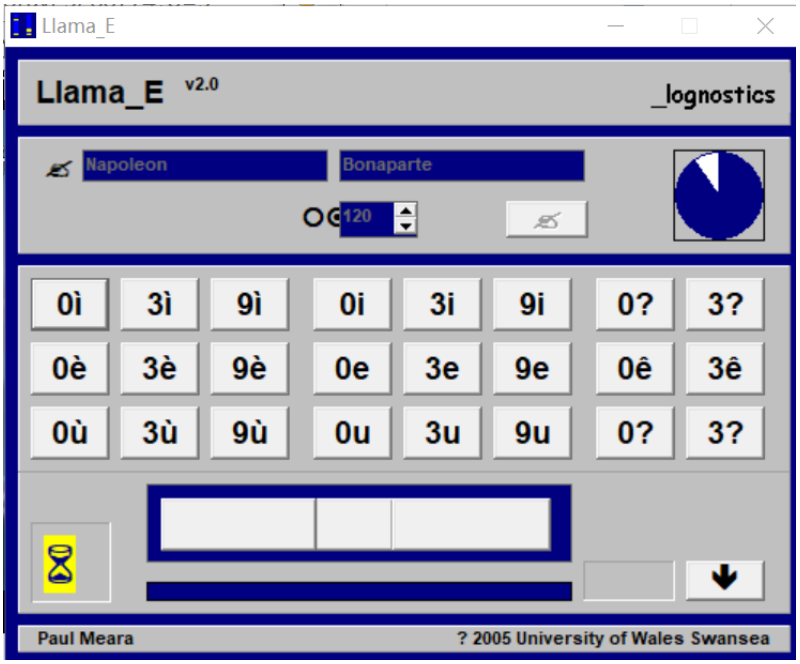

Figure 3.21 Screenshot of LLAMA E – Study phase

LLAMA E (Figure 3.21) was a test of sound-symbol association. In this test, the participants were asked to learn the alphabet of an artificial language and its pronunciation. The participants learnt the pronunciation of the alphabet by clicking on the letters, which were presented to them in the test panel. Two minutes were given for studying the alphabet and its pronunciation, and the participants could choose how they went about learning. After 2 minutes, the participants moved on to the testing phase. In the testing phase, some words were played to the participants, and two spellings of the words were also given. The participants were asked to select a spelling that they thought was correct. At the time of testing, the alphabet remained in the LLAMA E panel (Figure 3.22). Feedback to the answers of the participants was given in the form of a ding sound (correct answers) or a bleep sound (incorrect answers). At the end of the testing phase, the LLAMA E program automatically produced a test score.

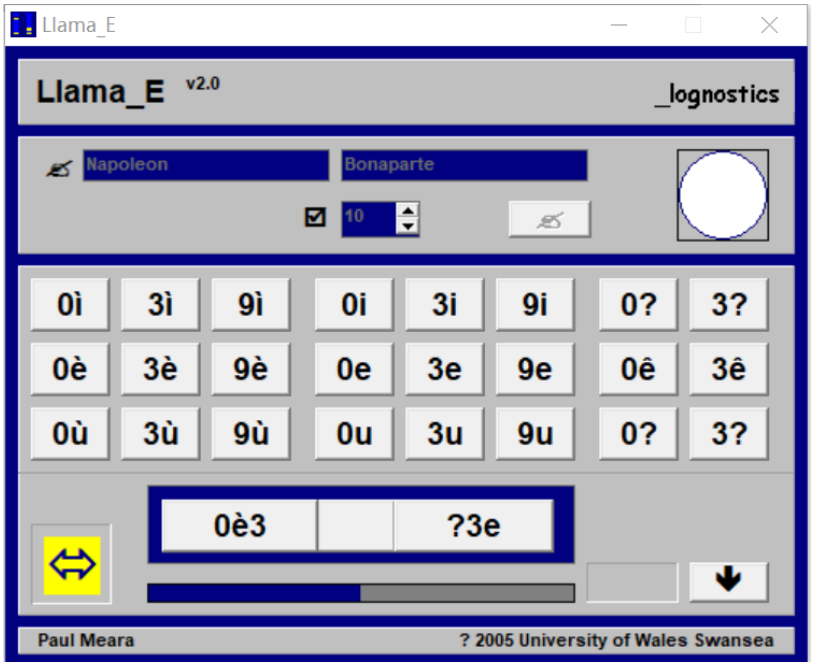

Figure 3.22 Screenshot of LLAMA E – Testing phase

LLAMA F was a test of grammatical inferencing ability. In this test, the participants were asked to study 20 picture stimuli and their corresponding sentences. The participants were given 5 minutes in the study phase, during which they learnt the picture-sentence pairings in their own way by clicking on the square buttons in the test panel (Figure 3.23).

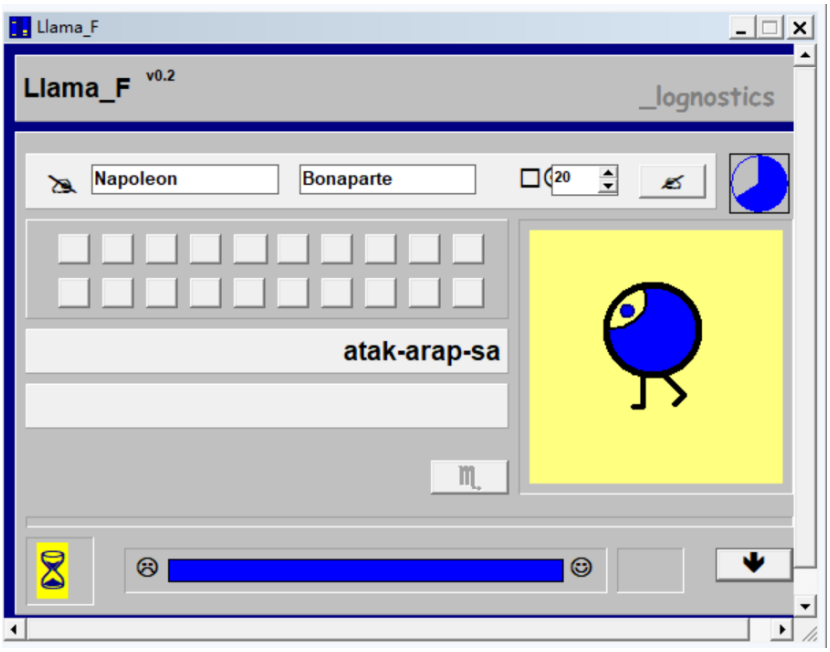

Figure 3.23 Screenshot of LLAMA F – Study phase

Then, in the testing phase, the participants were given some pictures and two sentences for each picture. They were asked to decide which sentence correctly described the given picture (Figure 3.24). Again, feedback was given in the form of a ding sound (correct answers) or a bleep sound (incorrect answers). At the end of the test, the LLAMA F program automatically produced a test score.

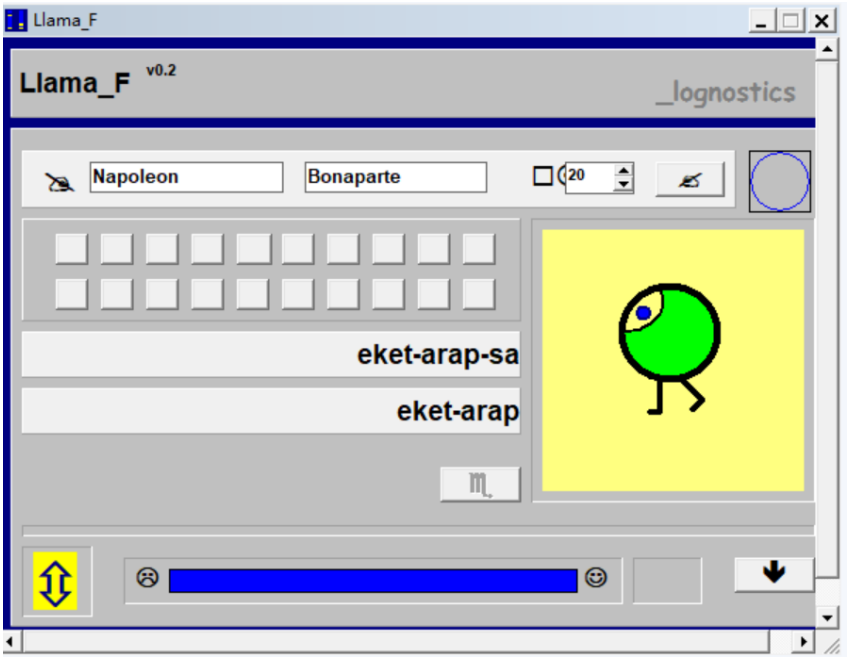

Figure 3.24 Screenshot of LLAMA E – Testing phase

3.4.6 The English learning experience questionnaire

A questionnaire was administered to the NNS in order to gather information about their demographic information, English learning experience and language use situation (Appendix 8). On the basis of R. Ellis et al. (2009), a questionnaire with four parts was developed.

In the first part, information AO, that is, the age that their learning of English had started in China, was gathered. Information about the length of study (i.e. years of study) of English in China before attending university, the frequency of their English classes at school, and the method of teaching (i.e. whether their teacher emphasised communication or grammar) was also gathered in this part.

In part 2, information about the learning experiences of the NNS in university was gathered. The researcher asked the participants to indicate whether they majored in English at university or not, and also to give information about the number of classes that they had at university.

Part 3 of the questionnaire was designed to gather information about the learning experience of the NNS in the L2 context. The participants were asked about their age of immersion (i.e. AoA), that is, the age that they immigrated to New Zealand. Questions regarding their learning experience after immigration, including learning English in a language class or attending New Zealand schools and universities were asked.

Lastly, part 4 of the questionnaire asked questions about the current language use of the NNS. They were also asked to report the years that they had spent working using English, and their length of residence in an English-speaking country. A list of activities that may reflect their daily language use context was provided to the participants, and they were asked to choose the ones that applied to them. For example, in this question, the participants were asked to indicate their language use situation at home, either by choosing "I speak more Chinese than English with my family at home in New Zealand." or "I speak more English than Chinese with my family at home in New Zealand."

The questionnaire was purposefully designed and administered in English. Since the participants were long-term residents in the L2 country and had been using English for work, it was expected that they would not have difficulties in answering a questionnaire in English. In other words, the implementation of this English questionnaire was intended as an informal test, which could be a tool to find out the participants with a very low level of proficiency. However, the researcher found that all the NNS participants were able to complete this English questionnaire on their own, which probably means that they were not learners with a low level of proficiency. Each NNS participant completed the questionnaire with pen and paper at the very end of the entire session.

3.5 Reliability of the test instruments

The reliability of all the test instruments was checked by calculating Cronbach's alpha reliability estimate (Cronbach, 1951). The general guideline of interpreting the alpha reliability estimate is: an alpha estimate bigger than .80 indicates good internal consistency, and an alpha bigger than .70 indicates an acceptable level of internal consistency. However, research has also shown that Cronbach's alpha can be inflated by a large number of samples (i.e. the number of items entered for the calculation of the alpha statistics), and deflated by a limited number of samples, and therefore these statistics should be viewed with caution (Cortina, 1993).

It was found that the tests of implicit and explicit knowledge had good reliability. The EI test was highly consistent internally, as indicated by α of .90. Since the EI test was marked by two raters, the extent to which these two raters agreed with each other was examined by calculating Cohen's kappa. The result of $\kappa = .79$, $p < .001$, indicated that inter-rater reliability was good. The internal consistency of the WordM test was also good, as indicated by high Cronbach's α of .93. Similarly, the TGJT showed high internal consistency with an α of .87. The UGJT, MKT and the cloze test received an α of .84, .76, and .77 respectively, all indicating an acceptable level of reliability.

The two tests of working memory similarly received a Cronbach's α reliability estimate bigger than .70, showing that they were internally consistent. The α statistics of the NonWord test and the SSpan test were .81 and .89, respectively. The inter-rater reliability of the NonWord test was also good, as indicated by the Cohen's kappa result, $\kappa = .78$, $p < .001$.

The LLAMA test received a relatively low α value of .623. This lower value might be a result of the limited number of subtests ($n = 4$) entered to calculate the statistics. As separate scores for each question in each sub-test of the LLAMA test were not available, only the scores of the four sub-tests were entered to calculate the reliability of this test. As a result, the α statistics appeared low.

Overall, based on these results, it was concluded that the measures used in the

present study had good reliability.

3.6 Pilot study

A pilot study that examined the feasibility of the instruments was conducted prior to the main study. The researcher also used this opportunity to practise test administration in preparation for the main study.

Six linguistics PhD students participated and completed the full test battery individually with the researcher. These participants were all NNS of English who speak Mandarin Chinese as their L1. The tests were administered following the exact procedures described in the previous sections. Special attention was paid to the tests programmed with DMDX version 5.1.4.2 (i.e. the SSpan, the EI, the TGJT, and the UGJT), and the output files of these tests were checked.

The researcher observed whether or not the 10-second time constraint imposed in the EI would be a further time pressure. It was found that there were cases when some participants were unable to finish their repetition within this time frame, so it was concluded that this 10-second time pressure would be implemented in the main study. Similarly, the 5-second time pressure in the TGJT caused some difficulties for the participants in the pilot study because there were cases when a judgement was not made within this time frame. Therefore, no changes were made to the allocated time in the EI and the TGJT in the main study.

The researcher also observed participants' performance in the NonWord test. In Zhao (2015), the decision was made to stop the test when the participant failed to repeat the non-words in three consecutive trials. Then the maximum number of correctly repeated non-words was taken as the final score of the participants. However, in the pilot study, the researcher found that the participants generally could remember trials with 4 to 5 words on average, but it was their ability to remember words in trials that exceeded 6 words that varied. Therefore, it was determined that all the participants were to finish all the trials in the main study. They were

encouraged to repeat as many words as possible irrespective of the total number of words in each trial. No other changes to the instruments were made in the main study.

The research methodology has been addressed so far in the thesis. In the next chapter, results will be presented.

CHAPTER FOUR

RESULTS

In this chapter, the results of the study are presented. The chapter starts with details of the preparation of data, and then proceeds to the descriptive statistics results of all the measures used in the study. Then the results of the inferential statistics are presented.

4.1 Data preparation

The entire dataset (N = 106) was separated into three parts, 1) language measures data, 2) language aptitude and working memory data and 3) questionnaire data. An initial round of data screening was conducted to identify the missing data in preparation for subsequent statistical analysis. After checking for data entry mistakes, the pattern of missing data was checked using Little's MCAR test (Hair, Black, Babin, & Anderson, 2014; Osborne, 2013).

For the six language measures, the following data was submitted for Little's MCAR test: the total scores of each test, and the sub-scores of the grammatical and ungrammatical stimuli sentences in the EI, the TGJT and the UGJT. Among this data, only a small amount of missing data, ranging from 0 to 4.7%, was found. The Little's MCAR test returned a Chi square value of .00, df = 131, p = 1.00, which suggested that the missing data was completely at random (MCAR). This kind of missing data does not cause bias to the observed data but might cause a reduction in sample size and therefore lead to reduced power (Osborne, 2013). In order to maintain the current sample size, the missing data was estimated using the expectation maximization method (EM) via IBM SPSS Statistics version 23.0. The EM process performed an iterative procedure in which the missing values were first estimated based on the current data matrix (expectation), and then checked as to whether the estimated values were most likely to appear (maximization). The use of

EM imputation has the advantage of preserving the relationship between the missing values and other variables and at the same time maintaining sample size (Tabachnick & Fidell, 2013). The imputed dataset also benefits further analyses such as factor analyses and multiple regressions.

The language aptitude and working memory data was complete, and therefore was retained after checking for data entry errors. Similarly, the questionnaire data was mostly complete, excluding some questions which were not applicable to some participants.

4.2 Descriptive statistics

Oral elicited imitation test

Following the advice of Loewen and Plonsky (2016), data normality was examined by employing multiple methods: 1) skewness and kurtosis statistics, and 2) the Kolmogorov-Smirnov test of normality. The criteria used for normal distribution were between ± 2 for skewness and ± 3 for kurtosis, and $p > .05$ in the Kolmogorov-Smirnov test (Field, 2009; Gravetter & Wallnau, 2014; Trochim & Donnelly, 2006).

It was found that the sum scores of the NS and the NNS in this test were normally distributed (NS: skewness = -.40, kurtosis = -.38; NNS: skewness = .12, kurtosis = -.39). The Kolmogorov-Smirnov tests of normality produced consistent results and also indicated normal distribution of the data of both groups. The test results of the scores of the NS were $D(20) = .16$, $p = .18$, and the NNS $D(86) = .09$, $p = .07$. No outliers were found within both groups. Table 4.1 below summarises the test results of the EI tests of the two groups, including the mean scores and standard deviations of each group. The total score for this test was 44.

Table 4.1 Descriptive statistics of the EI test

Group	NS (n = 20)		NNS (n = 86)	
	Mean	*SD*	Mean	*SD*
Sum score (max = 44)	39.80	2.53	26.63	7.40

It can be seen that there were between-group differences in the mean sum scores and the standard deviations. The NS group had a higher mean score (M = 39.8) with a smaller standard deviation (SD = 2.53) than the NNS (M = 26.63, SD = 7.40), which suggested better overall performance and smaller within group variation of the NS. An independent samples t-test was then conducted to statistically compare the performance of the two groups of participants. The results, $t(90.21) = 13.47$, $p < .001$, showed a significant between-group difference. The results of Levene's test for equality of variances, $F = 12.54$, $p = .001$, also indicated a significant difference between the standard deviations of the two groups.

Word monitoring test

In this test, the grammatical sensitivity index (GSI) was calculated by subtracting the mean word monitor times in milliseconds of the grammatical (MTG) from those of the ungrammatical sentences (MTUG). The skewness and kurtosis statistics of the GSI showed that the data of the NNS group was not normally distributed (NS: skewness = 1.03, kurtosis = 2.28; NNS: skewness = .62, kurtosis = 3.25). The Kolmogorov-Smirnov tests of normality indicated violations to normal distribution of the data of both groups. The tests results of the GSI of the NS were $D(20) = .26$, $p = .001$, and the NNS $D(86) = .11$, $p = .008$.

Since the data of both groups were not normally distributed, instead of using the standard z score method which relies on the mean and standard deviation to detect outliers, the more robust modified z score method was employed (Iglewicz & Hoaglin, 1993; Leys, Ley, Klein, Bernard, & Licata, 2013). In this method, the median and the median absolute deviation (MAD) were used to calculate the

modified *z* scores that are robust to non-normal distribution, and the threshold for outliers was values plus or minus 3.5 times the MAD (Iglewicz & Hoaglin, 1993). It was found that cases 24 and 15 in the NNS group and cases 10 and 13 in the NS group were outliers and thus the scores of these participants were removed.

Table 4.2 below shows the descriptive statistics of the word monitoring results of the NS and the NNS. In situations where the data does not comply with normal distribution, the median is a better representation of the central tendency of the data. Therefore, both the medians and the standard deviations are provided (Loewen & Plonsky, 2016).

Table 4.2 Descriptive statistics of the WordM test

Group	NS (n = 18)			NNS (n = 84)		
	Mean	Median	*SD*	Mean	Median	*SD*
MTG	567.06	552.88	137.74	885.86	753.39	405.90
MTUG	668.32	604.72	239.48	964.91	818.33	485.89
GSI	101.63	78.62	148.11	79.05	56.72	182.66

It is worth noticing that there were some negative GSI scores in both groups. This means that some participants took more time to monitor the target words in the grammatical sentences than they did for the ungrammatical ones. The overall mean GSI of the NNS was smaller than that of the NS. However, the results of the independent samples *t*-test, $t(102) = 1.37$, $p = .17$, showed that the GSI of the two groups were not significantly different. The standard deviations of the GSI of the two groups also did not differ significantly, as was shown by the result of Levene's test for equality of variances, $F = 1.12$, $p = .29$.

Timed grammaticality judgement test

In this test, each correct judgement of the grammaticality of the sentence was rewarded with 1 point. The total scores for the grammatical and the ungrammatical

sentences were both 34, and the total sum score was 68. The skewness and kurtosis statistics of the sum scores showed that the data of both groups were normally distributed (NS: skewness = -.55, kurtosis = 1.11; NNS: skewness = .65, kurtosis = -.28). The Kolmogorov-Smirnov tests of normality however, produced somewhat different results. The test result of the sum score of the NS group indicated a normal distribution ($D(20) = .17$, $p = .15$), but that of the NNS group a non-normal distribution ($D(86) = .13$, $p = .001$). No outliers were found in either group.

Table 4.3 shows the descriptive statistics of the judgement accuracy of the NS and the NNS, including the mean sum scores and the scores obtained for the grammatical (AccuracyG) and the ungrammatical sentences (AccuracyUG).

Table 4.3 Descriptive statistics of the TGJT

Group	NS (n = 20)		NNS (n = 86)	
	Mean	*SD*	Mean	*SD*
AccuracyG (max = 34)	31.00	1.21	29.43	3.21
AccuracyUG (max = 34)	29.15	2.87	15.32	6.86
Sum (max = 68)	60.15	2.74	44.75	7.17

It is apparent that the NS and the NNS differed in their mean sum scores (60.15 and 44.75) and in the standard deviations of mean sum scores (2.74 and 7.17). In order to test whether these between-group differences were statistically significant, an independent samples *t*-test and a Levene's test for equality of variances were conducted. The test result of the independent samples *t*-test, $t(81.58) = 15.60$, $p < .001$, showed that the two groups differed significantly in mean sum scores. Similarly, the result of Levene's test, $F = 15.09$, $p < .001$, indicated that the variances of the two groups were statistically different.

The results of two independent samples *t*-tests showed that the differences between the two groups were also evident in their judgement accuracies of the grammatical ($t(82.89) = 3.53$, $p = .001$) and the ungrammatical sentences ($t(74.60) = 13$, $p < .001$). The NNS obtained a mean score of 29.15 on the grammatical

sentences, but only 15.32 on the ungrammatical ones. In contrast, the NS performed equally well on both types of sentences (*M* = 31 and 29.15 respectively). This indicated that the grammaticality of the stimuli sentences played a role in influencing the judgement accuracy of the NNS, but not the NS (see Section 4.4.1).

Untimed grammaticality judgement test

As in the timed grammaticality judgement test, each correct judgement of the grammaticality of a sentence was rewarded 1 point. The total scores for the grammatical and the ungrammatical sentences were both 34, and the total sum score was 68. The skewness and kurtosis statistics of the sum scores in this test showed that the data of the NS group was not normally distributed (skewness = -2.11, kurtosis = 6.15, but that of the NNS group was normally distributed (skewness = -.95, kurtosis = .77). However, the Kolmogorov-Smirnov tests of normality showed that the data of both groups violated normal distribution, as indicated by *p* values below the significance level of .05 for the NS group *(D*(20) = .23, *p* = .008) and the NNS group (*D*(86) = .11, *p* = .018). Table 4.4 summarises the descriptive statistics of this test.

Table 4.4 Descriptive statistics of the UGJT

Group	NS (*n* = 19)			NNS (*n* = 85)		
	Mean	Median	*SD*	Mean	Median	*SD*
AccuracyG (max = 34)	33.15	34	1.87	30.08	31	4.03
AccuracyUG (max = 34)	31.45	32	1.82	29.10	30	4.39
Sum (max = 68)	63.80	65	2.91	58.33	59	6.63

Visual inspections of the histograms found that the data of both groups were negatively skewed, which indicated that most participants obtained relatively high scores in this test. Since unlimited time was allowed for judgement, the NS were expected to score near-ceiling and thus the negatively skewed data was not a surprise. The NNS also demonstrated a similar distribution pattern with only a few scores

below 50 (n = 9). However, some outliers were found in both groups. In the NS group, participant 8 had a modified z score of -4.95 and was identified as an outlier. Similarly, participant 19 in the NNS group (modified z = -3.07) was identified, and therefore the scores of these participants were removed.

Similarly, an independent samples t-test and a Levene's test for equality of variances were conducted to assess the degree of difference in the judgement accuracy of the two groups. It was found that there were statistical differences in the mean sum scores (t(95.72) = 7.17, p < .001) and the standard deviations (F = 15.24, p < .001) between the NS and the NNS.

Metalinguistic knowledge test

There were three parts in this test, and each correct answer was awarded 1 point. The skewness and kurtosis statistics of the sum scores in this test showed that the data of both groups were normally distributed (NS: skewness = .23, kurtosis =-.88; NNS: skewness = -.53, kurtosis = -.76). However, somewhat different results were indicated by the Kolmogorov-Smirnov tests of normality. The NS group received a Kolmogorov-Smirnov statistics D(20) =.12, p = .20, from which a normal distribution was indicated. The NNS group received a D(86) = .13, p = .001, and this significant p value indicated violations to normal distribution. The histogram of the NNS group was then checked and a negatively skewed distribution pattern was observed. Most NNS participants scored around 30 out of 39, and this caused the asymmetric distribution of the data. No outliers were found in both groups.

Table 4.5 Descriptive statistics of the MKT

Group	NS (n = 20)		NNS (n = 86)	
	Mean	*SD*	Mean	*SD*
Part 1 (max = 17)	9.50	2.95	10.86	2.99
Part 2 (max = 18)	8.65	4.80	11.42	4.86
Part 3 (max = 4)	1.50	1.43	1.77	1.13
Sum (max = 39)	19.65	8.18	24.05	7.54

Table 4.5 summarises the descriptive statistics of this test, including the descriptive statistics of each individual part. The total score for each part was 17, 18, and 4 respectively; and the total sum score was 39. The NNS group outperformed the NS group in this test. An independent samples *t*-test was conducted to examine the between group difference, and the result $t(104) = -2.32$, $p = .02$, indicated a statistical difference in the mean sum scores. However, the between-group difference in the variances of the sum score were not statistical, as indicated by $F = .10$, $p = .76$ in Levene's test for equality of variances.

Cloze test

The cloze test had 20 questions, and the total score was 41. The skewness and kurtosis statistics of the total score in this test showed that the data of both groups were normally distributed (NS: skewness = -.19, kurtosis = -.62; NNS: skewness = .05, kurtosis = -.57). The Kolmogorov-Smirnov tests of normality also indicated normal data distribution of the NS group ($D(20) = .13$, $p = .20$) and the NNS group ($D(86) = .09$, $p = .07$). No outliers were found. Table 4.6 is a summary of the scores obtained by the participants in this test.

Table 4.6 Descriptive statistics of the cloze test

Group	NS (n = 20)		NNS (n = 86)	
	Mean	*SD*	Mean	*SD*
Sum (max=41)	27.3	7.02	21.82	7.21

This test caused the NS some difficulty and the NNS even more so. To the surprise of the researcher, several participants in the NS group found this test very difficult and even gave up halfway. Since this was the last test in the entire test battery, it was possible that the relatively long testing period (1 hour for the NS, and 2 hours for the NNS) had exhausted the participants at that point. Both the NS and the NNS group scored relatively low in this test (the NS group only scored 70% of the total, and the NNS only scored 53%). The result of the independent samples

t-test, $t(104) = 3.08$, $p = .003$, again showed that the two groups differed significantly in their mean sum scores. The large standard deviations of the two groups (7.02 and 7.21) indicated big within-group variances in test performance.

Non-word repetition test

The NonWord test played a total of 24 trials and required the NNS to recall the non-words they heard in each trial. Following the percentage method, that is, the number of correctly repeated non-words were divided by the total number of non-words in each trial and then averaged out across all trials (Conway et al., 2005; Friedman & Miyake, 2005; Thompson & Sykes, 2010), the maximum possible score for this test was 1. The skewness and kurtosis statistics of this data (skewness = -.36, kurtosis = -.01) indicated normal distribution, which was inconsistent with the results from the Kolmogorov-Smirnov test statistics $D(85) = .10$, $p = .03$.

Within the NNS group, participant 30 ($z = -3.58$) was identified as an outlier. In retrospect, this participant was an elderly lady who relied on hearing aids at the time of participation in the research. She encountered difficulties when doing this task and found the stimuli difficult to hear. Consequently, she obtained a score that was distinctively lower than the other participants. Considering the fact that this task relied heavily on listening perception, this score was removed. The descriptive statistics are shown in Table 4.7.

Table 4.7 Descriptive statistics of the NonWord test

Group	NNS ($n = 85$)	
	Mean	*SD*
Sum (max = 1)	.46	.09

Listening sentence span test

The listening SSpan test was also scored using the percentage method, and the maximum possible score was 1. Again, only the NNS completed this test. The skewness and kurtosis statistics of the sum score showed that the data was normally

distributed (skewness = -.13, kurtosis = -.31). The Kolmogorov-Smirnov test of normality also provided evidence of a normal distribution, $D(85) = .06$, $p = .20$. However, the score of participant 68 ($z = -3.33$) was found as an outlier and thus was removed. The descriptive statistics are shown in Table 4.8.

Table 4.8 Descriptive statistics of the SSpan test

Group	NNS (n = 85)	
	Mean	*SD*
Sum (max = 1)	.57	.13

LLAMA test

The LLAMA test had four sub-tests and the scores were automatically provided by the LLAMA program. Each sub-test had a total score of 100. It was found that the standard deviations in the four sub-tests were rather large, which indicated that there were large variations in the scores of the NNS in this test. It is also possible that the large standard deviations were a result of data non-normality. Table 4.9 shows the descriptive statistics of this test.

Table 4.9 Descriptive statistics of the LLAMA test

Group	NNS (n = 86)		
	Mean	Median	*SD*
B (max = 100)	51.86	50	18.50
D (max = 100)	26.69	25	14.90
E (max = 100)	43.60	40	25.89
F (max = 100)	49.65	50	24.71
Sum (max = 400)	171.80	170	60.17

Judging from Kolmogorov-Smirnov test statistics of $D(86) = .10$, $p = .02$; $D(86) = .12$, $p = .004$, $D(86) = .13$, $p = .001$ for sub-tests D, E, and F, respectively, it was confirmed that the scores of these three sub-tests were not normally distributed.

However, the skewness and kurtosis of the total sum score (max = 400) of the data (skewness = .26, kurtosis = -.15) indicated a normal distribution of the sum scores, in line with the Kolmogorov-Smirnov test statistics $D(86) = .06$, $p = .20$.

4.3 Validating the tests of implicit and explicit knowledge

The present study set out to investigate the implicit and explicit knowledge of the participants with the aim of understanding the long-term outcomes of second language acquisition. Hence valid measures of these two types of knowledge are of crucial importance; without these sound inferential statistical analyses could not be done and convincing conclusions could not be made. Therefore, before answering any research questions, the first step was to validate the six language tests and examine whether the hypothesised type of knowledge was reliably measured. In other words, the extent to which the six language tests served as separate measures of implicit and explicit knowledge was investigated first. For this purpose, correlation analyses and factor analyses were conducted.

It should be noted that only NNS data was used in the factor analyses in the present study and this decision was made based on some statistical considerations. For the purpose of instrument validation, the common practice in factor analysis is to use single group data (Brown, 2015). This is because factor analysis depends on the variance-covariance matrix of the data, which in turn depends on how participants are sampled from the target population. The NS and the NNS participants in the present study came from two heterogeneous populations, and therefore their data should not be congregated in factor analysis. If the data of these two groups of participants were combined into a single dataset, its variance-covariance matrix would neither be representative of the NS group, nor the NNS group. Considering the fact that the NNS group was of primary research interest, the data of the NS group was excluded in the factor analyses.

The exclusion of the data of the NS group in factor analyses does not

necessarily mean that NS's data would not be useful for validating the tests of L2 knowledge. The NS participants in the present study formed a comparison group. These 20 NS were asked to complete the same set of language measures to allow for a baseline set of data to be obtained, against which the acquisition outcomes of the NNS could be compared. In other words, the primary aim of the study was to examine the outcomes of L2 acquisition of the NNS, but the analysis of the NNS alone may not be meaningful without looking at the difference between them and the NS. The differences between the NS and NNS, in fact, would be some evidence that attests to the construct validity of the language measures (see Sections 5.1.1 and 5.1.2). However, for the reasons stated in the previous paragraph, the NS data was not included in the factor analyses.

4.3.1 Correlation analysis

Among the six tests, the hypothesised measures of implicit L2 knowledge were the EI, the TGJT, the WordM test and the cloze test. The UGJT and the MKT were the hypothesised measures of explicit L2 knowledge. Table 4.10 shows the Pearson correlation coefficients matrix of the six language tests (n = 86). These coefficients not only represent the covariance between variables, but also indicate the magnitude (i.e. the effect size) of the correlations. The criteria for determining the effect sizes are that r values close to .25 denote a small effect size; .40 a medium effect size; and .60 a large effect size (Plonsky & Oswald, 2014, p. 889).

It can be seen from Table 4.10 that the WordM did not have significant correlations with any other language tests. The correlation coefficients were low, and the effect sizes were small (r ranged between .05 to .15). Another test that had relatively weak correlations with the others was the MKT. It only had a significant correlation with the UGJT (r = .56, p < .01), which was another hypothesised measure of explicit knowledge. This significant correlation and the close to large effect size provided some evidence of the construct validity of the MKT as a

measure of explicit knowledge. In addition, its low and non-significant correlations with the hypothesised measures of implicit knowledge (*r* ranged from .07 to .18) were also an indication that this test measured a different type of knowledge from what was measured by the EI and the TGJT.

Table 4.10 Correlation matrix for the language measures of the NNS group (*n* = 86)

	EI	TGJT	WordM	UGJT	MKT	Cloze
EI	__	.70 **	.05	.64 **	.18	.53 **
TGJT		__	.15	.49 **	.07	.33 **
WordM			__	.12	.08	.12
UGJT				__	.56 **	.63 **
MKT					__	.36 **
Cloze						__

** $p < .01$

The high correlation between the EI and the TGJT ($r = .70$, $p < .01$) showed that the knowledge that was accessed by these two tests was similar. However, both of these tests also had high correlations with the hypothesised measure of explicit knowledge UGJT (EI and UGJT, $r = .64$, $p < .01$; TGJT and UGJT $r = .49$, $p < .01$), which implied some possible overlap in what was measured by these tests.

The cloze test had significant correlations with all the other tests except for the WordM test. The highest correlation was with the UGJT ($r = .63$, $p < .01$), which not only indicated a large effect size, but also a possibly stronger link between this test and explicit L2 knowledge. However, it also correlated largely and significantly with the EI ($r = .53$, $p < .01$). Meanwhile, correlations of the cloze test with the TGJT and the MKT were also observed ($r = .33$ and .36 respectively, $p < .01$), indicating medium effect sizes.

To sum up, Pearson's correlation coefficients revealed a complex pattern. The failure to find significant correlations between the WordM test and the other hypothesised tests of implicit knowledge, and between the MKT and the other

hypothesised tests of explicit knowledge, indicated the need for further in-depth investigations. The results of the correlation coefficients of the cloze test also called for further analysis.

4.3.2 Exploratory factor analysis

Considering the complex pattern of the correlations among the six language measures, and the fact that several measures (i.e. the EI, the TGJT, and the WordM) were modified from the original tests (see R. Ellis et al., 2009), an exploratory factor analysis (EFA) was conducted to determine the number of latent constructs that these measures tapped.

The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) statistics of .69 indicated adequate sampling (Kaiser, 1974). Bartlett's test of sphericity $\chi^2(15) = 183.62$, $p < .01$, indicated that the correlations between the tests were sufficient for a factor analysis. Factors were extracted using the principal axis factoring method based on the correlation matrix (i.e. the standardised covariance matrix) using the Direct Oblimin rotation method. The principal axis factoring method, which does not require distributional assumptions of the data, was chosen as a suitable method for an EFA (Brown, 2015). The Direct Oblimin rotation method belongs to the family of oblique rotation, which allows for correlations between factors (Brown, 2015; Pituch & Stevens, 2015). This method was chosen because of the hypothesised possible structure of the factors (i.e. implicit and explicit knowledge are probably correlated). The factor extraction followed the Kaiser-Guttman rule, that factors with eigenvalues greater than 1 were retained (Pituch & Stevens, 2015). The scree plot was also checked as a supplementary factor extraction criterion (Figure 4.1).

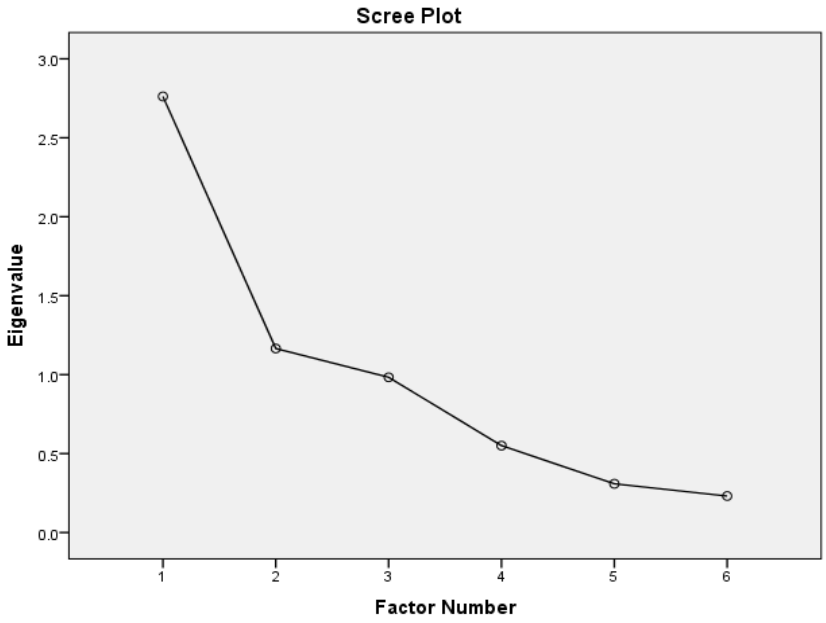

Figure 4.1 Scree plot of the exploratory factor analyis of language measures

The results of the analysis showed a two-factor solution that in combination accounted for 66.19% of the total variance. Considering the relatively small sample size of the analysis (< 100), factor loadings greater than .50 (corresponds to 25% variance explained) were retained for further interpretation (Hair et al., 2014). Table 4.11 presents the results of the analysis after factor rotation.

Table 4.11 Exploratory factor analysis results of the language measures

	Factor		Communalities
	1	2	
EI	.90		.82
TGJT	.77		.61
WordM			.02
UGJT	.69	.89	.92
MKT		.67	.44
Cloze	.51	.59	.47

Note. Factor loadings smaller than .50 are not shown.

The EFA results of the six language measures revealed some interesting findings. To begin with, the two-factor solution yielded was consistent with the hypothesised constructs of implicit and explicit knowledge. Tests that were intended as measures of implicit knowledge were the EI, the TGJT, the WordM test and the cloze test. While the former two tests loaded on the first factor with high loadings (.90 and .77 respectively), the WordM test did not have enough loadings to be retained in either factor. The communalities results echoed the factor loading situation for this factor, which showed the extraction communalities of .82, .61, and .02 for the first three tests respectively. This indicated that while the hypothesised factor of implicit knowledge explained most of the variance of the EI and the TGJT, it hardly explained any variance of the WordM test. In other words, the type of knowledge that the WordM test measured was possibly irrelevant to implicit knowledge. It was probably irrelevant to explicit knowledge also since it did not receive sufficient loading for the second factor (i.e. explicit knowledge).

While the cloze test, as a hypothesised measure of implicit knowledge, had a loading of .51 on the hypothesised factor implicit knowledge, it had an even higher loading of .59 on the second factor, namely explicit knowledge. The two hypothesised measures of explicit knowledge, the UGJT and the MKT had loadings of .89 and .67 respectively on this factor. However, it should be noted that the UGJT also had a substantive loading on the first factor (factor loadings = .69). The cross loadings of the cloze test and the UGJT suggested that both factors explained part of the variance of these two tests, indicating that both types of knowledge were probably used in the testing process. However, the higher loadings on the second factor of these two tests suggested that these tests had stronger connections with explicit knowledge. In comparison, the MKT only loaded on the second factor, reflecting its role as a more valid measure of explicit knowledge that is distinct from implicit knowledge.

This two-factor solution provided a good fit of the data in general, as was indicated by the 66.19% total variance explained and the 13% non-redundant

residuals. For a good fit model, the cut-off is to have less than 50% of the non-redundant residuals with absolute values greater than .05 (Brown, 2015). These results implied that most of the six language measures performed as desired. However, the instances of the cross-loadings of the UGJT and the cloze test raised some concerns. The cloze test, as an intended measure of implicit knowledge, surprisingly had higher loadings on the hypothesised factor of explicit knowledge. Therefore, it is too early to confirm the validity of these measures and further analysis is necessary.

4.3.3 Confirmatory factor analysis

The results of the EFA implied that most of the language measures were consistent with the theoretical hypotheses. These include: 1) the TGJT and the EI are primarily measures of implicit knowledge, and 2) the UGJT and the MKT are primarily measures of explicit knowledge, although the UGJT might also elicit the use of implicit knowledge. However, the other hypotheses were not borne out, including: 1) The WordM test primarily measures implicit knowledge, and 2) the cloze test measures implicit knowledge. Discrepancies between the hypotheses and the existing results analysis suggest that the validity of these language measures needs further analysis. Therefore, a confirmatory factor analysis (CFA) was conducted based on the results of the exploratory factor analysis.

Prior to the construction and evaluation of a structural model, the Mahalanobis distance of the six language measures was calculated in order to check for multivariate outliers, followed by a calculation of its probability level using the χ^2 distribution. According to Tabachnick and Fidell (2013), cases with Mahalanobis distance greater than the χ^2 critical value at the probability of .001 are likely to be multivariate outliers. The critical value of χ^2 at $df = 6$ is 22.458. The calculated Mahalanobis distance values ranged from .47 to 16.49 and the corresponding p values ranged from .998 to .011. It was therefore concluded that no multivariate

outliers were found.

The multivariate skewness and kurtosis of the six language measures were checked by conducting Mardia's multivariate normality test (Mardia, 1970; 1974) using R (Korkmaz, Goksuluk, & Zararsiz, 2014; R Core Team, 2017). The test results returned a Mardia's skewness of 7.01, $p < .001$ and a kurtosis of 49.31, $p = .054$, indicating a multivariate non-normal distribution. The multivariate Shapiro-Wilk normality test (Royston, 1983) was also conducted to examine data normality using R (Jarek, 2012; R Core Team, 2017), and the test result $W = .093$, $p < .001$ again indicated violations to normal distribution.

The structural model CFA1 (Figure 4.2) in line with the EFA results was constructed using the Lavaan package (Rosseel, 2012) under the R environment (R Core Team, 2017). The UGJT and the cloze test were specified as indicators of the factor explicit knowledge to which higher loadings were observed in the EFA. Due to the multivariate non-normality of the data, the maximum likelihood with robust standard errors and χ^2 (i.e. the MLM estimator) was employed for model estimation (Bentler & Chou, 1987; Brown, 2015). The data of the NNS was then entered to fit this model.

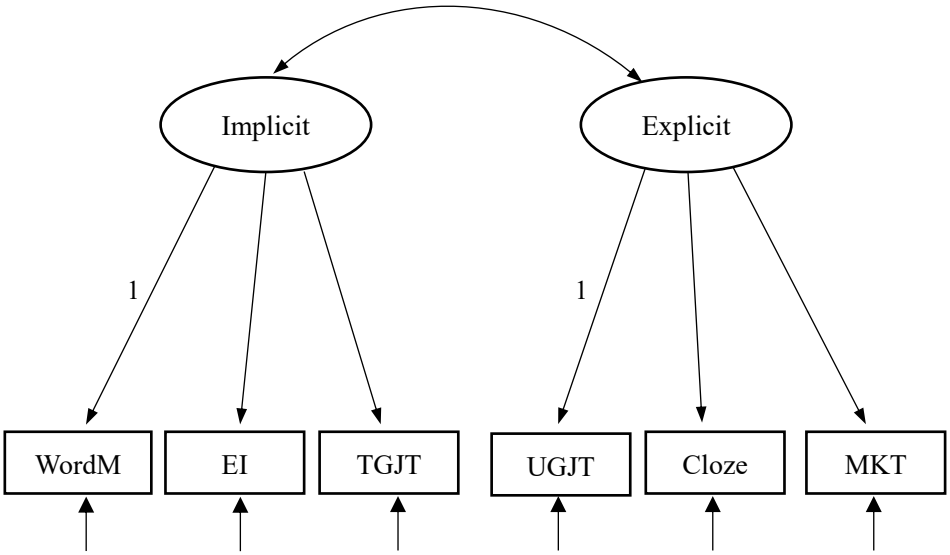

Figure 4.2 Structural model CFA1

Unfortunately, the Chi square test result of model CFA1, χ^2 (8) = 18.37, p = .02, indicated poor model fit. The p value of less than .05 suggested that this model was not properly defined and did not capture some of the underlying structure of the dataset (Brown, 2006; Byrne, 2010). Three types of goodness-of-fit indices were also examined, including the absolute fit index standardized root mean square residual (SRMR), and two comparative fit indices comparative fit index (CFI) and Tucker-Lewis index (TLI). The widely reported parsimony fit index root mean square error of approximation (RMSEA) was also checked, but it was interpreted with caution due to its unstable performance in models with small degrees of freedom (Kenny, Kaniskan, & McCoach, 2015). The threshold for determining model fit were: .06 ≤ SRMR < .08 indicating acceptable fit and SRMR < .06 indicating good fit (L. Hu & Bentler, 1999); RMSEA < .05 indicating good fit and RMSEA < .08 indicating acceptable fit for models with a sample size of below 250 (Fabrigar & Wegener, 2012; MacCallum, Browne, & Sugawara, 1996); CFI > .95 and TLI > .95 indicating good model fit (L. Hu & Bentler, 1999). The results of these fit indices of model CFA1 are reported in Table 4.12, which also indicate an unsatisfactory model fit.

Table 4.12 Goodness-of-fit indices of model CFA1

GoF type	Absolute	Parsimony	Comparative	
	SRMR	RMSEA (90% CI)	CFI	TLI
Statistics	.07	.13 (.05, .20)	.94	.89

The modification indices of model CFA1 were checked in order to improve model fit. Indices of 3.84 or greater (correspond to Chi-square at p < .05, 1 *df*) suggest an overall improvement to model fit (Brown, 2015; Hair et al., 2014). However, an important premise in using the modification indices is that it should be based on sufficient and compelling theoretical or methodological considerations because some modification indices can be nonsensical (Byrne, 2016). In addition,

the standardised residuals in the covariance matrix of this model were also checked to identify problematic model specifications. Standardised residuals that are equal to or greater than the absolute value of 1.96 (corresponding to a significance level of .05) are generally regarded as significant (Hair et al., 2014). Positive large values indicate that the covariance between the indicators is not sufficiently accounted for, while large negative values indicate an overestimation of the covariance relationship.

Given the information provided by the modification indices, the standard residuals, the results of the exploratory factor analysis, and some methodological considerations together, the following two adjustments were proposed: 1) a possible error covariance between the UGJT and the MKT test, and 2) the removal of the redundant WordM test as an indicator of the factor of implicit knowledge due to its very low loading ($\lambda = .10$). Accordingly, model CFA2 was constructed (Figure 4.3) and fitted using the MLM estimator.

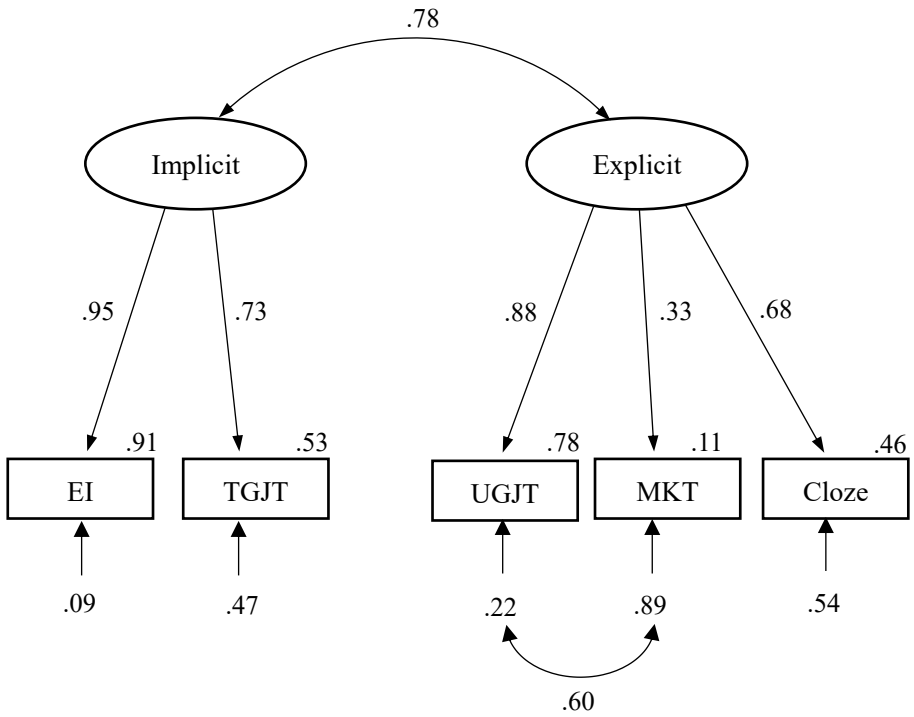

Figure 4.3 Modified factor model CFA2

The overall model fit was significantly improved as indicated by $\chi^2(3) = 7.59$, $p = .06$. The goodness-of-fit indices results are presented in Table 4.13. The

parameter estimates, the error variances estimates and the covariance estimates are shown in Tables 4.14, 4.15 and 4.16, respectively.

Table 4.13 Goodness-of-fit indices of model CFA2

GoF type	Absolute	Parsimony	Comparative	
	SRMR	RMSEA (90% CI)	CFI	TLI
Statistics	.048	.134 (.000, .262)	.975	.915

Table 4.14 Parameter estimates of model CFA2

Factor	Indicator	B	*SE*	*z*	*p*	β	R^2
IK	EI	1	--	--	--	.95	.91
IK	TGJT	.74	.12	6.16	.00	.73	.53
EK	UGJT	1	--	--	--	.88	.77
EK	MKT	.45	.14	3.35	.00	.33	.11
EK	Cloze	.88	.15	6.02	.00	.68	.46

Table 4.15 Error variances estimates of model CFA2

Factor	Indicator	B	*SE*	*z*	*p*	β
IK	EI	4.77	5.01	.95	.34	.09
IK	TGJT	23.86	5.57	4.28	.00	.47
EK	UGJT	8.73	3.70	2.20	.03	.22
EK	MKT	51.43	7.28	7.07	.00	.89
EK	Cloze	27.42	5.00	5.48	.00	.54

Table 4.16 Covariance estimates of model CFA2

Factor/Indicator	Factor/Indicator	B	*SE*	*z*	*p*	β
IK	EK	30.06	5.38	5.59	.000	.78
UGJT	MKT	12.68	4.27	2.97	.003	.60

According to model CFA2, both tests of implicit knowledge had an acceptable

level of factor loading. The hypothesised construct of implicit knowledge explained 91% of variance in the data of the EI, and 53% of variance in the TGJT. These results were in line with the results of the previous exploratory factor analysis.

In comparison, a different pattern emerged for the three tests of explicit knowledge. Among these three, only the UGJT obtained a sufficient factor loading of .88 that explained 77% of its variance. The factor loading of the cloze test was .68, which was slightly below the cut off value of .70 and explained 46% of the variance. The factor loading of the MKT was relatively low ($\lambda = .33$), which indicated that only a small amount of the variance in this test were explained by the hypothesised factor explicit knowledge. The significant error covariance between the MKT and the UGJT ($\delta = .60$, $z = 2.97$, $p = .003$, see Table 4.16) suggested that the unexplained covariance of these two was due to some exogenous factor. The relatively lower factor loadings of the UGJT, the cloze test, and the MKT were partly in line with previous EFA results.

The correlation of the hypothesised two factors (i.e. implicit L2 knowledge and explicit L2 knowledge) was .78, $p < .01$; a value that is close to the cut off value of .85 as an indication of discriminant validity (Brown, 2015). Therefore, to further examine the discriminant validity of these two factors, a constrained model that fixed the between-factor correlation to 1 was constructed and compared to model CFA2 (Satorra & Bentler, 2001). The Chi square difference test result $\chi^2_{diff} = 105.42$, $p < .001$ indicated that the correlation between the two factors was significantly different from 1, lending support to the discriminant validity of these two factors. Therefore, it was concluded that implicit and explicit L2 knowledge were correlated, yet were different.

An alternative model which specified the cloze test as a measure of implicit knowledge was also tested. This model specification was in line with the hypothesis that the cloze test measures implicit knowledge. However, this model did not converge and therefore it was concluded that in the present study the cloze test was primarily linked to explicit L2 knowledge.

Theoretically, implicit and explicit knowledge represents L2 language knowledge (R. Ellis, 2008b). Model CFA2 showed that implicit and explicit knowledge were correlated, but the nature of this relationship was unaddressed. In order to examine the relationship between the two factors, a higher-order confirmatory factor model CFA3 was built (Figure 4.4).

In model CFA3, a higher-level factor L2 knowledge was specified to account for the variance of implicit and explicit knowledge. In order to achieve model identification, the higher order factor loadings of implicit and explicit knowledge were specified to have equal weights (Byrne, 2016; Hair et al., 2014). Since the possession of both implicit and explicit knowledge was most likely the case for the majority of the NNS participants in the present study, this specification was considered appropriate. The Chi square test of the higher order model CFA3 indicated a good model, $\chi^2(3) = 7.71$, $p = .052$. The goodness-of-fit indices are shown in Table 4.17.

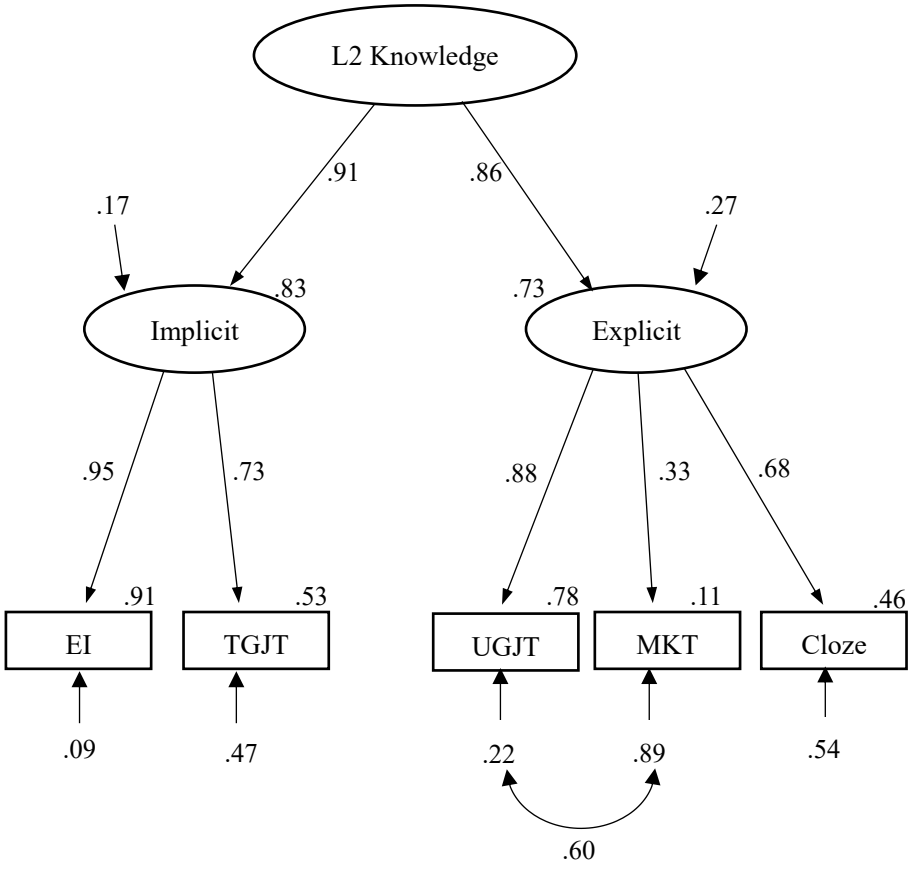

Figure 4.4 Higher-order factor model CFA3

Table 4.17 Goodness-of-fit indices of model CFA3

GoF type	Absolute	Parsimony	Comparative	
	SRMR	RMSEA (90% CI)	CFI	TLI
Statistics	.056	.140 (0.00, 0.26)	.973	.910

These indices showed that the overall fit of this higher order model was acceptable. In fact, the parameter estimates of the indicators of implicit and explicit knowledge remained the same as model CFA2. Implicit and explicit knowledge were identified as two significant manifest factors of L2 knowledge, with factor loadings of .91 and .86 respectively. These correspond to the explanation of 83% of variance in implicit knowledge and 73% of variance in explicit knowledge. The higher factor loading of implicit knowledge indicated a stronger link of this type of knowledge with L2 knowledge.

4.4 The effect of test stimulus on grammaticality judgement accuracy

Understanding the effect of test stimulus, that is, the sentence type (i.e. grammatical or ungrammatical) of the stimuli sentences has been a part of the analysis of grammaticality judgement tests. Some previous research has indicated that the grammatical and the ungrammatical sentences in the TGJT and the UGJT play different roles and measure different constructs (Loewen, 2009). In fact, scores of the grammatical and ungrammatical items in these tests have been analysed separately in factor analyses previously (e.g. Gutierrez, 2013). Results have suggested that the grammatical and ungrammatical sentences constitute different measures of implicit and explicit knowledge irrespective of test condition (i.e. timed or not). One approach that directly tests whether the grammatical and the ungrammatical sentences measure the same construct is by using factor analysis, where the scores of the grammatical and the ungrammatical sentences are used as indicators. However, for this method to be used in the present study, the data of a

minimum number of 140 participants (calculated according to the 10:1 ratio of cases to free parameters in the model) would have been required (Bentler & Chou, 1987). The participant number of 86 failed to meet this requirement, so the influence of test stimuli was analysed independently from the previous factor analyses.

Statistically speaking, the data obtained from the TGJT and the UGJT are categorical in nature since participants are required to make correct/incorrect judgements. The common way of analysing this kind of data is by using aggregated scores which are then transformed into proportions, followed by analysis of variance (ANOVA). For example, two sum scores of the grammatical and ungrammatical sentences would be calculated in a grammaticality judgement test, and the means and variances of these two sum scores would then be compared. However, it has been pointed out that ANOVA is not the best way to analyse categorical data. Specifically, the use of aggregated data not only violates the underlying assumption of ANOVA, but may also lead to spurious results (Jaeger, 2008). Alternatively, logistic regression analysis and mixed logit models have been proposed as better statistical methods for the analysis of categorical data (Agresti, 2002; Baayen, Davidson, & Bates, 2008; Harrell, 2015; Bates, Maechler, Bolker, & Walker, 2015; Jaeger, 2008). Mixed logit models have the advantage of controlling for the random effects of test materials and participants (i.e. test materials and participants are drawn from a much larger pool of materials and populations), and thus provide a more accurate estimation of the influence of sentence grammaticality on judgement accuracy.

To understand the role that sentence type played in the TGJT and the UGJT in the present study, several generalised mixed models were constructed and compared. These mixed models were similar to logistic regression analyses, which predict the log odds of binary responses based on one or more continuous and/or categorical variables. In addition, the mixed models included an estimation of the effects of randomly selected participants and test materials. By doing so, mixed model analyses not only enable the investigation of the factor of research interests (i.e. test stimulus in the present study), but also statistically estimate the magnitude of

participant specific variations.

4.4.1 Sentence type in the TGJT

The data of the NS and the NNS of the TGJT were analysed separately. Three logistic mixed models M0NS, M1NS and M2NS were constructed using the data of the NS group with the lme4 package (Bates, Machler, & Bolker, 2017) under the R environment (R Core Team, 2017).

In model M0NS, participants' response accuracy was modelled as a function of the fixed effect of sentence type and the random effect of test item. In model M1NS, an additional simple subject (i.e. participant) random effect was specified on the basis of model M0NS. This model explained the influence that sentence type had on response accuracy, assuming that participants and stimuli sentences were chosen randomly. In model M2NS, a participant-specific random effect of sentence type was added on the basis of model M1NS. This model included all the fixed and random effects in the previous model, and also allowed for the influence of sentence type to be different for each subject.

M0NS: Accuracy ~ SentenceType + (1| Item)

M1NS: Accuracy ~ SentenceType + (1| Subject) + (1| Item)

M2NS: Accuracy ~ SentenceType + (1+ SentenceType| Subject) + (1| Item)

Table 4.18 shows the parameter estimates of the fixed effects in the above three models (n = 1360), and Table 4.19 shows the random effects estimates.

Table 4.18
Parameter estimates of the fixed effects in models M0NS, M1NS and M2NS

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
M0NS	Intercept	3.54	.42	8.5	.00 ***
	Sentence	-.87	.52	-1.67	.09
	Type=Ungrammatical				
M1NS	Intercept	3.58	0.43	8.38	.00 ***
	Sentence	-.88	0.53	-1.67	.10
	Type=Ungrammatical				
M2NS	Intercept	3.60	0.43	8.30	.00 ***
	Sentence	-.75	0.59	-1.28	.20
	Type=Ungrammatical				

*** $p < .001$

Table 4.19
Parameter estimates of the random effects in models M0NS, M1NS and M2NS

Model	Predictor	Variances	*SD*	Correlation
M0NS	Item (Intercept)	3.08	1.76	
M1NS	Subject (Intercept)	.09	.29	
	Item (Intercept)	3.16	1.78	
M2NS	Subject (Intercept)	.05	.22	
	Subject (ST=Ungrammatical)	.89	.94	-1
	Item (Intercept)	3.34	1.83	

These three models were then compared by conducting a likelihood ratio test. The results are shown in Table 4.20.

Table 4.20 Comparison of models M0NS, M1NS and M2NS

	Df	AIC	BIC	logLik	Chisq Chi	*df*	Pr(>Chisq)
M0NS	3	782.33	797.98	-388.17			
M1NS	4	782.88	803.74	-387.44	1.448	1	.229
M2NS	6	775.47	806.76	-381.74	11.412	2	.003 **

*** $p < .001$

The likelihood ratio test statistics $X^2_{(1)} = 1.45$, $p = .229$ showed that model M2NS was not statistically different from model M0NS. However, the likelihood ratio test statistics between models M1NS and M2NS, $X^2_{(2)} = 11.41$, $p < .05$, provided evidence in favour of the latter model. The change of 2 degrees of freedom resulted in a rather big change in X^2 value of 11.41, associated with a significant p value (< .05). This indicated that the second model better explained the data (Bates, 2010; Faraway, 2016). The advantage of model M2NS over M1NS was also reflected in the smaller Akaike Information Criterion (AIC) value (Bates et al., 2017). Therefore, model M2NS was selected as the final model.

In model M2NS, the intercept represented the log odds of making a correct judgement when the stimuli sentence was grammatical. The intercept was taken as the baseline of comparison, to which the influence of sentence type was compared. The coefficient of the sentence type (ungrammatical) represented the change in the log odds of making a correct judgement when the stimuli sentence changed from grammatical to ungrammatical. The positive coefficient of the intercept 3.58 indicated that overall the probability of making a correct judgement by the NS when the stimuli sentence was grammatical was very high, equivalent to 97% in probability. However, once they were presented with an ungrammatical stimuli sentence, there was a change in the log odds of making a correct judgement by -.75. In other words, the change of stimuli sentences from grammatical to ungrammatical caused a decreased probability of making a correct judgement by approximately 32%. Although this change in probability might seem a significant figure, it should be noted that the Wald test did not produce a significant result. The Wald test was used to describe how distant the coefficient estimates are from zero in terms of their standard errors, with above 3 absolute z values and smaller than .05 p values suggesting statistically significant parameter estimates (Bates et al., 2017; Jaeger, 2008). This means that the fixed effect of sentence type did not influence the performance of the NS substantially.

On the basis of the fixed effect, the random effects of subject and stimuli

sentences can be interpreted. As the same set of sentences was used with every participant, only a random intercept (represented by 1|Item) was specified for stimuli sentences. This meant that whatever the effects of sentence type might be, it was the same for all the stimuli sentences. In contrast, the specification of the subject random effect also included a random slope (represented by 1 + SentenceType | Subject), which captured the inter-participant variability in terms of how they were affected by sentence type. In other words, the underlying assumption of this specification was that sentence type might have different effects on different participants.

The usefulness of the random slope assigned to each participant in model M2NS was reflected in the likelihood ratio test, in which a significant lowering of the AIC statistics was observed (AIC decreased from 782.88 to 775.47). The relatively small intercept of the random effects estimation indicated that there was no big difference (s^2 = .05) in the overall performance of the NS. However, there was a difference (s^2 = .89) in terms of how they were influenced by sentence type. That is to say, although sentence type did not influence the performance of the NS at a group level, there were slight differences in the way that each participant was influenced. Nonetheless, this difference was minimal in comparison to the other random effect of stimuli sentences (s^2 = 3.34).

Similarly, three logit mixed models M0NNS, M1NNS and M2NNS were constructed and compared using the data of the NNS. These three models were specified in exactly the same way as the three models of the NS group. Table 4.21 shows the parameter estimates of the fixed effects in these models (n = 5780), and Table 4.22 shows the random effects estimates. The likelihood ratio test results of model comparison are shown in Table 4.23.

M0NNS: Accuracy ~ SentenceType + (1| Item)

M1NNS: Accuracy ~ SentenceType + (1| Subject) + (1| Item)

M2NNS: Accuracy ~ SentenceType + (1+ SentenceType| Subject) + (1| Item)

Table 4.21
Parameter estimates of the fixed effects in models M0NNS, M1NNS and M2NNS

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
M0NNS	Intercept	1.99	.13	14.78	.00 ***
	Sentence Type=Ungrammatical	-2.21	.19	-11.89	.00 ***
M1NNS	Intercept	2.13	.16	13.30	.00 ***
	Sentence Type=Ungrammatical	-2.36	.20	-11.83	.00 ***
M2NNS	Intercept	2.17	.17	12.45	.00 ***
	Sentence Type=Ungrammatical	-2.42	.26	-9.18	.00 ***

*** $p < .001$

Table 4.22
Parameter estimates of the random effects in models M0NNS, M1NNS and M2NNS

Model	Predictor	Variances	*SD*	Correlation
M0NNS	Item (Intercept)	.50	.70	
M1NNS	Subject (Intercept)	.41	.64	
	Item (Intercept)	.57	.76	
M2NNS	Subject (Intercept)	.46	.68	
	Subject (ST=Ungrammatical)	1.90	1.38	-.67
	Item (Intercept)	.69	0.83	

Table 4.23 Comparison of models M0NNS, M1NNS and M2NNS

	Df	AIC	BIC	logLik	Chisq Chi	*df*	Pr(>Chisq)
M0NNS	3	5923.2	5943.2	-2958.6			
M1NNS	4	5732.6	5759.2	-2862.3	192.58	1	.000 ***
M2NNS	6	5529.3	5569.3	-2758.7	207.27	2	.000 ***

*** $p < .001$

It can be seen from Table 4.23 that the models with both the participant and sentence random effects (i.e. M1NNS and M2NNS) outperformed the model

(M0NNS) with only the stimuli sentence random effect. The test statistics of model M1NNS in comparison to model M0NNS, $X^2_{(1)} = 192.58$, $p < .001$, indicated that the addition of the participant random effect significantly improved the overall model fit (Faraway, 2016). The addition of the random slope to the participant random effect further improved the model fit, as was indicated by the likelihood ratio test statistics $X^2_{(2)} = 207.27$, $p < .001$. This model also had the lowest AIC and Bayesian information criterion (BIC) values, which indicated its superiority. Therefore, model M2NNS was selected as the best model.

For the NNS, sentence type was found to have had a significant influence on their performance (Wald $z = -9.18$, $p < .001$). The intercept indicated the log odds of making a correct judgement when the stimuli sentence was grammatical (Intercept = 2.17), which corresponds to 89.73% in probability. The negative coefficient of sentence type = ungrammatical (-2.42) indicated that when the stimuli sentence changed from grammatical to ungrammatical, the change in the log odds of making a correct judgement decreased by 2.42. That is to say, the odds of making a correct judgement decreased multiplicatively by exp(-2.42) = .089 when the stimuli sentences were ungrammatical, which corresponds to a 91% decrease in the accuracy of their judgement.

The random effect of stimuli sentences was rather small, as was indicated by $s^2 = .69$. In comparison, there was a big difference in the way that sentence type influenced each participant ($s^2 = 1.90$). This indicated that in addition to the general negative influence of sentence type on judgement accuracy, it had a different influence on each participant. In other words, the NNS were not uniformly influenced by sentence type, and there were differences in terms of how individuals were affected.

To conclude, the grammaticality of the stimuli sentence influenced the performance of the NS and the NNS differently. While the ungrammatical sentences did not cause much decrease in the judgement accuracy of the NS, it led to a significant decrease in the accuracy of the NNS.

In order to examine how big the difference was between the NS and the NNS, the data of both groups were aggregated into one dataset, and then two mixed logit models M1 and M2 were constructed and compared. A new variable "Group" was also added to differentiate the data of the two groups. In these models, group and sentence type were specified as fixed effects, and subject and stimuli sentences as random effects.

M1: Accuracy ~ SentenceType + Group + (1 |Subject) + (1|Item)

M2: Accuracy ~SentenceType + Group + (1+SentenceType|Subject) + (1|Item)

Table 4.24 presents the results of the parameter estimates of the fixed effects of these two models. The parameter estimates of the random effects are shown in Table 4.25, and the likelihood ratio test results of model comparison are shown in Table 4.26.

Table 4.24 Parameter estimates of the fixed effects in models M1 and M2

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
M1	Intercept	3.75	.19	19.33	.00 ***
	Sentence	-2.17	.20	-10.68	.00 ***
	Type=Ungrammatical	-1.74	.11	-15.23	.00 ***
	GroupNNS				
M2	Intercept	3.80	.20	18.74	.00 ***
	Sentence	-2.21	.25	-8.73	.00 ***
	Type=Ungrammatical	-1.77	.12	-15.09	.00 ***
	GroupNNS				

*** $p < .001$

Table 4.25 Parameter estimates of the random effects in models M1 and M2

Model	Predictor	Variances	*SD*	Correlation
M1	Subject (Intercept)	.33	.57	
	Item (Intercept)	.62	.79	
M2	Subject (Intercept)	.35	.59	
	Subject (ST=Ungrammatical)	1.48	1.22	-.66
	Item (Intercept)	.70	.84	

Table 4.26 Comparison of models M1 and M2

	Df	AIC	BIC	logLik	Chisq Chi	*df*	Pr(>Chisq)
M1	5	6662.2	6696.6	-3326.1			
M2	7	6488.2	6536.3	-3237.1	178.01	2	.000 ***

*** $p < .001$

The likelihood ratio test statistics $X^2_{(2)} = 178.01$, $p < .001$, indicated the superiority of model M2 over model M1. This model showed that both sentence type and group had significant influences. The intercept 3.80 demonstrated that the log odds of making a correct judgement for the grammatically correct stimuli sentences of the NS was equivalent to a 97.8% in probability. This result was very similar to the result obtained in model M2NS. The coefficient estimate -2.17 of sentence type suggested that the log odds of making a correct judgement when the stimuli sentences were ungrammatical decreased by 2.21 in comparison to grammatical stimuli sentences. That is to say, there was an 89% [(exp (-2.21)-1) *100%] decrease in terms of probability. The coefficient estimate -1.77 of group indicated the log odds of making a correct judgement by the NNS were 1.77 times lower than that of the NS, which corresponds to 83% lower in probability.

Figure 4.5 visually depicts these differences in the predicted probabilities of judgement accuracy according to sentence type by groups.

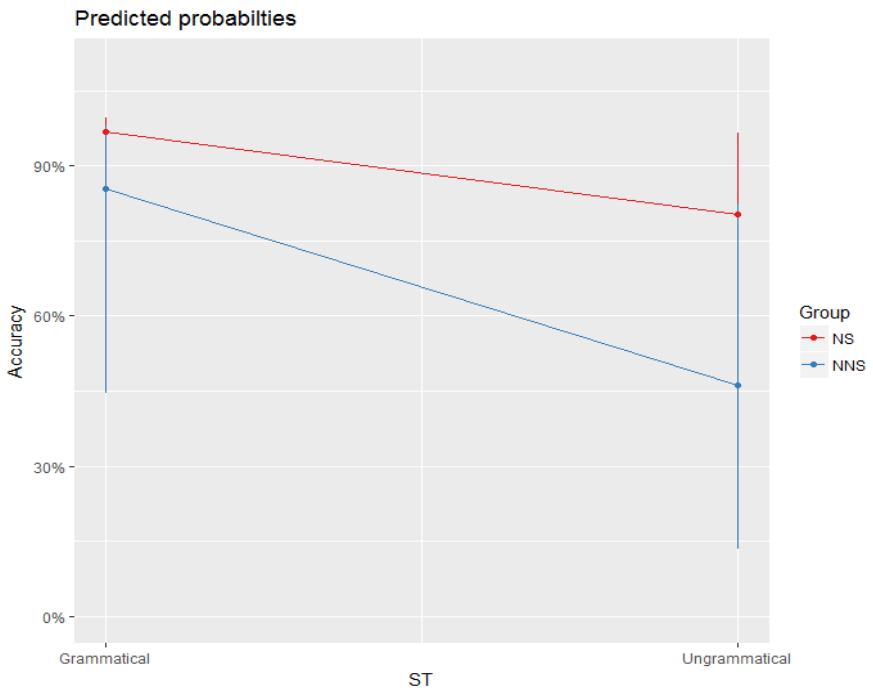

Figure 4.5 Predicted probabilities of TGJT judgement accuracy

It can be seen from Figure 4.5 that the difference between the NS and the NNS groups was substantial, and the negative effects of sentence type and group were evident. When given grammatical stimuli sentences, the possibility of making an accurate judgement by the NS was nearly 95% and around 85% for the NNS. When encountering ungrammatical sentences, this possibility lowered to around 80% for the NS, while it dropped dramatically to only around 45% for the NNS. To put it another way, the slopes of the two horizontal lines indicated the extent that sentence type influenced judgement accuracy, and the relatively steep slope of the NNS group demonstrated a very significant negative influence.

The vertical lines were the corresponding 95% confidence intervals of the predicted probabilities. The NS group had narrower confidence intervals, indicating higher accuracy of the predicted probabilities, in contrast to the observably wider ones of the NNS group. Some overlaps of the confidence intervals were observed between these groups, and these indicated the possibility that some accurate NNS performed at a standard that fell within the NS range.

4.4.2 Sentence type in the UGJT

To investigate the role of sentence type in the UGJT, the data of the NS and the NNS were analysed separately where sentence type was specified as a fixed effect predictor. Using the data of the NS, the following three logit mixed models M3NS, M4NS, and M5NS were constructed:

M3NS: Accuracy ~ SentenceType + (1| Item)
M4NS: Accuracy ~ SentenceType + (1| Subject) + (1| Item)
M5NS: Accuracy ~ SentenceType + (1+ SentenceType| Subject) + (1| Item)

Table 4.27 shows the parameter estimates of the fixed effects in these models (n = 1360), and Table 4.28 shows the random effects estimates. The likelihood ratio test results of model comparison are shown in Table 4.29.

Table 4.27
Parameter estimates of the fixed effects in models M3NS, M4NS and M5NS

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
M3NS	Intercept	3.95	.41	9.72	.00 ***
	Sentence Type=Ungrammatical	-.84	.47	-1.79	.07
M4NS	Intercept	4.15	.45	9.21	.00 ***
	Sentence Type=Ungrammatical	-.88	.49	-1.80	.07
M5NS	Intercept	4.41	.54	8.16	.00 ***
	Sentence Type=Ungrammatical	-1.16	.58	-2.01	.04 *

*** $p < .001$ * $p < .05$

Table 4.28
Parameter estimates of the random effects in models M3NS, M4NS and M5NS

Model	Predictor	Variances	*SD*	Correlation
M3NS	Item (Intercept)	1.92	1.39	
M4NS	Subject (Intercept)	.31	.56	
	Item (Intercept)	2.11	1.45	
M5NS	Subject (Intercept)	.84	.92	
	Subject (ST=Ungrammatical)	.55	.74	-.87
	Item (Intercept)	2.25	1.50	

Table 4.29 Comparison of models M3NS, M4NS and M5NS

	Df	AIC	BIC	logLik	Chisq Chi	*df*	Pr(>Chisq)
M3NS	3	571.47	587.12	-282.74			
M4NS	4	564.79	585.65	-287.39	8.69	1	.003 **
M5NS	6	566.53	597.82	-277.27	2.25	2	.324

** $p < .05$

The likelihood ratio test results showed that model M4NS was the best model. In comparison to model M3NS with only the simple random effects of item, M4NS, with the simple random effects of both item and subject was a superior model ($X^2_{(1)} = 8.69$, $p < .005$). The more complex M5NS showed no improvement to M4NS ($X^2_{(2)} = 2.25$, $p = .32$).

In model M4NS, the significant intercept coefficient (4.15) represented the log odds of making correct judgements when the stimuli sentences were grammatical. This was equivalent to 98.4% in probability and showed that the accuracy rate was very high. The coefficient of sentence type = ungrammatical showed a change in the log odds of making correct judgements when stimuli sentences changed from grammatical to ungrammatical. The negative coefficient -.88 indicated a negative influence of sentence type; however it was not significant (Wald's $z = -1.80$, $p = .07$). That is to say, although the ungrammatical stimuli sentences caused a slight decrease

in the judgement accuracy of the NS, this decrease was not statistically significant. The estimation of random effects in M4NS showed a higher estimate of s^2 for stimuli sentences than subject, which indicated greater individual variability among the stimuli sentences, rather than among participants.

The same model construction and comparison procedure was conducted to examine the influence of sentence type on judgement accuracy using the data of the NNS (n = 5640). These results are shown in Tables 4.30, 4.31, and 4.32.

M3NNS: Accuracy ~ SentenceType + (1| Item)

M4NNS: Accuracy ~ SentenceType + (1| Subject) + (1| Item)

M5NNS: Accuracy ~ SentenceType + (1+ SentenceType| Subject) + (1| Item)

Table 4.30
Parameter estimates of the fixed effects in models M3NNS, M4NNS, and M5NNS

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
M3NNS	Intercept	2.03	.14	14.20	.00 ***
	Sentence Type=Ungrammatical	-.14	.20	-.68	.50
M4NNS	Intercept	2.27	.18	12.57	.00 ***
	Sentence Type=Ungrammatical	-.15	.20	-.68	.49
M5NNS	Intercept	2.40	.20	11.82	.00 ***
	Sentence Type=Ungrammatical	-.16	.27	-.58	.56

*** $p < .001$

Table 4.31
Parameter estimates of the random effects in models M3NNS, M4NNS, and M5NNS

Model	Predictor	Variances	*SD*	Correlation
M3NNS	Item (Intercept)	.55	.74	
M4NNS	Subject (Intercept)	.66	.81	
	Item (Intercept)	.66	.81	
M5NNS	Subject (Intercept)	1.08	1.04	
	Subject (ST=Ungrammatical)	1.44	1.20	-.58
	Item (Intercept)	.73	.86	

Table 4.32 Comparison of models M3NNS, M4NNS and M5NNS

	Df	AIC	BIC	logLik	Chisq Chi	*df*	Pr(>Chisq)
M3NNS	3	4436.1	4456.1	-2215.1			
M4NNS	4	4203.7	4230.2	-2097.8	234.49	1	.000 ***
M5NNS	6	4113.3	4153.1	-2050.7	94.34	2	.000 ***

*** $p < .001$

It can be seen from the likelihood ratio test results (Table 4.32) that model M4NNS was better than model M3NNS ($X^2_{(1)} = 234.49$, $p < .001$), and model M5NNS was even better than model M4NNS ($X^2_{(2)} = 94.34$, $p < .001$). Compared to model M3NNS which only included a sentence random effect, the inclusion of a simple subject random effect (1|Subject) in model M5NNS brought about significant improvement. This indicated that the variations among participants contributed largely to the model, which was also reflected by the $s^2 = .66$ in the random effects estimation. The addition of a subject random slope according to sentence type (1 + SentenceType | Subject) also proved significant, indicating its importance in the model. This means that on top of the overall negative influence of sentence type, each participant was affected differently. The variation caused by sentence type among the NNS group was big, as indicated by the $s^2 = 1.44$ in the random effects

estimation of model M5NNS.

Again, the analysis results indicated that sentence type had different influences on the two groups in the UGJT. Another two mixed logit models M3 and M4 were constructed to further examine the differences between the two groups. In these models, group and sentence type were specified as two fixed effects, and subject and stimuli sentences as random effects.

M3: Accuracy ~ SentenceType + Group + (1 |Subject) + (1|Item)

M4: Accuracy ~SentenceType + Group + (1+SentenceType|Subject) + (1|Item)

Table 4.33 presents the results of the parameter estimates of the fixed effects of these two models. The parameter estimates of the random effects are shown in table 4.34.

Table 4.33 Parameter estimates of the fixed effects in models M3 and M4

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
M3	Intercept	3.26	.22	14.86	.00 ***
	Sentence	-.22	.22	-1.02	.31
	Type=Ungrammatical	-.98	.14	-6.98	.00 ***
	GroupNNS				
M4	Intercept	3.42	.24	14.30	.00 ***
	Sentence	-.26	.26	-0.99	.32
	Type=Ungrammatical	-1.00	.14	-7.07	.00 ***
	GroupNNS				

*** $p < .001$

Table 4.34 Parameter estimates of the random effects in models M3 and M4

Model	Predictor	Variances	*SD*	Correlation
M3	Subject (Intercept)	.54	.74	
	Item (Intercept)	.69	.83	
M4	Subject (Intercept)	.98	.99	
	Subject (ST=Ungrammatical)	1.22	1.10	-0.64
	Item (Intercept)	.75	.87	

The likelihood ratio test results of model comparisons are shown in Table 4.35. The likelihood ratio test statistics $X^2_{(2)} = 83.54$, $p < .001$ indicated the superiority of model M4 over model M3. This model showed that group significantly influenced judgement accuracy, while sentence type did not. The coefficient estimate of group -1.00 indicated that the log odds of making a correct judgement by the NNS were 1.00 times lower than those of the NS, which corresponds to 26.8% lower in probability. The influence of group and sentence type on this test is shown in Figure 4.6.

Table 4.35 Comparison of models M3 and M4

	Df	AIC	BIC	logLik	Chisq Chi	*df*	Pr(>Chisq)
M3	5	4790.0	4824.3	-2390.0			
M4	7	4710.5	4758.4	-2348.2	83.54	2	.000 ***

*** $p < .001$

It can be seen from Figure 4.6 that the difference between the NS and the NNS was not as significant as model M2. The non-significant influence of sentence type was shown by the similar estimates of accuracy probability. The difference between the NS and NNS group was significant, as shown by the gap between the red line (NS) and the blue line (NNS). The overlaid vertical lines of the corresponding 95% confidence intervals of the predicted probabilities also indicated differences between the two groups. The NS group had narrower confidence intervals, indicating higher

accuracy of the predicted probabilities.

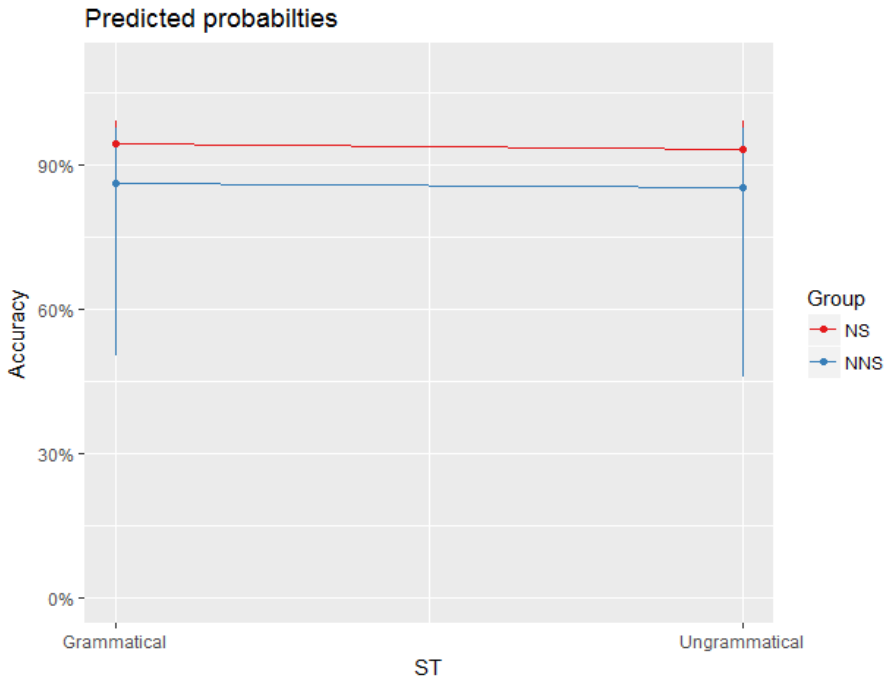

Figure 4.6 Predicted probabilities of UGJT judgement accuracy

4.5 Confidence ratings and sources of judgement in the UGJT

In the UGJT, participants were asked to provide confidence ratings. They were asked to do so because the relationship between confidence ratings and judgement accuracy was of interest. To examine the relationship between the confidence ratings of participants and their judgement accuracy, several mixed logit models were constructed.

In mixed models, with an ordered variable such as certainty level on the scale of 1 to 5, the method of orthogonal polynomial coding is used (Bates et al., 2012). Being a type of contrast coding, this method tests the trend or shape of the relationship between an ordered independent variable and the dependent variable. For a variable with 5 levels, 4 possible relationships are produced and tested, including the linear, quadratic, cubic and quartic relationships. The estimated coefficients and the associated Wald test results indicate the significance of each

relationship.

In the present study, there was no doubt that the NS were fairly confident with their judgement, so it was expected that the confidence ratings of these participants would not influence their judgement accuracy. This was confirmed by the results of a logit mixed model CertaintyNS (Table 4.36), in which judgement accuracy was modelled as a function of the fixed effects of certainty level and two simple random effects of subject and stimuli sentences. The results indicated no relationship between the certainty level of the native speakers and judgement accuracy (Wald *z* values below ± 3, $p > .05$).

CertaintyNS: Accuracy ~ Certainty + (1|Subject) + (1|Item)

Table 4.36 Parameter estimates of the fixed effects of model CertaintyNS

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
CertaintyNS	Intercept	-.82	144.64	-.01	.995
	Certainty. L	11.88	457.39	.03	.979
	Certainty. Q	-7.03	386.57	-.02	.985
	Certainty. C	3.67	228.70	.02	.987
	Certainty^4	-1.14	86.44	-.01	.989

In contrast, it was expected that the NNS would have more variation in their level of certainty, and thus a relationship could possibly be observed. Therefore, using the data of the NNS, two logit mixed models, CertaintyNNS and CertaintyNNS1 ($n = 5638$), were constructed and compared.

CertaintyNNS: Accuracy ~ Certainty + (1| Subject) + (1| Item)

CertaintyNNS1: Accuracy ~ Certainty + (1 + Certainty | Subject) + (1| Item)

In model CertaintyNNS, judgement accuracy was modelled as a function of the

fixed effect of level of certainty and two simple random effects of subject and stimuli sentences. In model CertaintyNNS1, a random slope of certainty was specified to each subject in addition to all the effects in model CertaintyNNS, which allowed for variation among participants in terms of certainty.

Table 4.37 presents the estimates of the fixed effects of these two models. The parameter estimates of the random effects are shown in Table 4.38, and the model comparison results are shown in Table 4.39.

Table 4.37
Parameter estimates of the fixed effects of models CertainyNNS and CertaintyNNS1

Model	Predictor	Coefficient	*SE*	Wald *z*	*p*
CertaintyNNS	Intercept	1.12	.17	6.68	.00 ***
	Certainty. L	2.07	.32	6.58	.00 ***
	Certainty. Q	.14	.28	0.52	.61
	Certainty. C	.34	.24	1.42	.16
	Certainty^4	.06	.17	0.37	.72
CertaintyNNS1	Intercept	.97	.17	5.543	.00 ***
	Certainty. L	2.65	.42	6.37	.00 ***
	Certainty. Q	-.06	.33	-0.18	.86
	Certainty. C	.46	.27	1.74	.08
	Certainty^4	-.01	.18	-0.05	.96

*** $p < .001$

Table 4.38
Parameter estimates of the random effects of models CertainyNNS, CertaintyNNS1

Model	Predictor	Variances	*SD*	Correlation
CertaintyNNS	Subject (Intercept)	.49	.70	
	Item (Intercept)	.57	.76	
CertaintyNNS1	Subject (Intercept)	.30	.55	
	Certainty. L	2.33	1.53	-.23
	Item (Intercept)	.59	.77	

Table 4.39 Comparison of models CertainyNNS and CertaintyNNS1

	Df	AIC	BIC	logLik	Chisq Chi	*df*	Pr(>Chisq)
CertaintyNNS	7	4030.0	4076.5	-2008.0			
CertaintyNNS1	21	4017.8	4157.2	-1987.9	40.22	14	.000 ***

*** $p < .001$

The likelihood ratio test statistics $X^2_{(14)} = 40.22$, $p < .05$, indicated that the more complex model CertaintyNNS1 was a better model. The parameter estimates of this model indicated that certainty level had a linear relationship with judgement accuracy (Certainty. L's $z = 6.39$, $p < .001$), and the positive coefficient 2.65 indicated a positive relationship between increase in level of certainty and judgement accuracy. In other words, when participants made judgements with higher levels of confidence, it was more likely that their judgements were accurate. Figure 4.7 demonstrates this relationship.

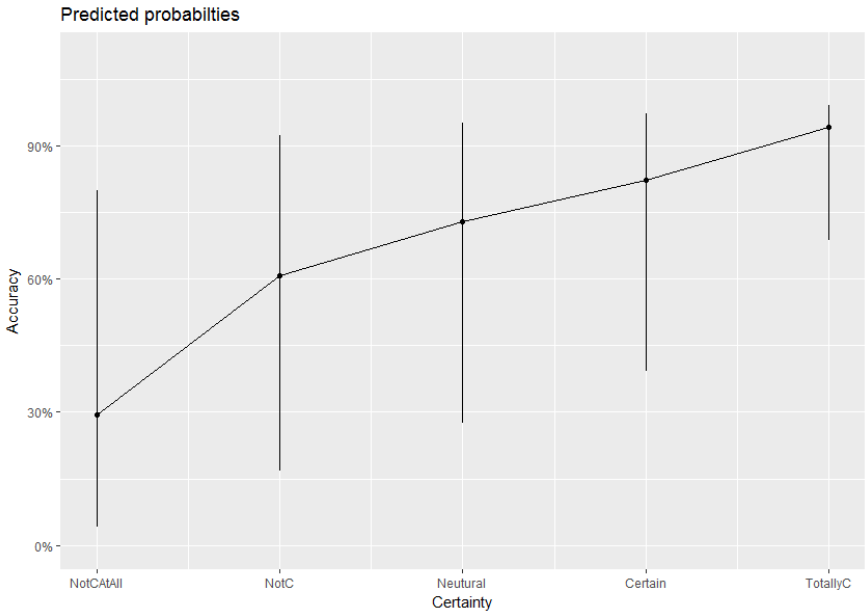

Figure 4.7 Predicted probabilities of judgement accuracy according to certainty

Note: NotCAtAll: not certain at all; NotC: not certain; TotallyC: totally certain.[2]

[2] The linear relationship of certainty and accuracy was in logit form, while the above plot was based on probabilities. Therefore, the somewhat curvy relationship depicted above was a result of the transformation from logit to probability.

The vertical lines associated with each point estimate of the predicted probability of accuracy were the corresponding confidence intervals. The relatively wide confidence intervals were in line with the estimation of the random effects, which showed big individual variations in confidence levels (as indicated by certainty. L, s^2= 2.33, *SD* = 1.53).

Each participant made a choice of whether they relied on rule, intuition or both rule and intuition to make judgements in the UGJT. Of the 84 NNS, 61 participants who said that they relied on rule (72.6%), 16 reported that they relied on intuition (19%), and 11 reported that they relied both on grammatical rule and intuition (8.3%). The mean scores of those relied on rule, on intuition, and on both rule and intuition were 58.90 (*SD* = 6.10), 57.94 (*SD* = 6.30), and 57.29 (*SD* = 8.88), respectively. An ANOVA was then conducted to examine if there were significant differences in UGJT performances between those who relied on rule, on intuition and on rule and intuition. The analysis results $F(2, 81) = .30$, $p = .74$ indicated that there was no difference in the test performance of the UGJT in relation to reports of how judgement decisions were made.

4.6 Comparison of implicit and explicit knowledge

On the basis of the results of the CFA (see Section 4.3.3), weighted average sum scores of the hypothesised factors implicit L2 knowledge (IK) and explicit L2 knowledge (EK) were calculated. With this method, the scores of each test contribute to sum scores with different weights (i.e. the factor loadings), that is, the tests that had higher factor loadings contributed more than those that had lower factor loadings. Since model CFA2 (see Section 4.3.3) has been verified as a well-fit model, the calculation of this kind of weighted sum scores of the hypothesised factors was considered the most appropriate (DiStefano, Zhu, & Mindrila, 2009).

The statistics of skewness and kurtosis of these weighted scores ranged between -1.11 to .59, from which little deviation from normal distribution was indicated

(Field, 2009; Gravetter & Wallnau, 2017). The results of Kolmogorov-Smirnov tests of normality of the two groups also indicated normal distribution of the variables named IK $D(19) = .12$, $p = .20$ and $D(85) = .09$, $p = .19$; and EK, $D(19) = .11$, $p = .20$, and $D(85) = .08$, $p = .20$. No outliers were found within both groups. The descriptive statistics are presented in Table 4.40.

Table 4.40 Descriptive statistics of IK and EK factor scores

Group	NS (n = 19)		NNS (n = 85)	
	Mean	*SD*	Mean	*SD*
IK (max = 45.72)	40.86	1.83	28.98	5.67
EK (max = 33.82)	27.37	2.55	24.92	3.63

To facilitate comparisons between IK and EK, these scores were then divided by the maximum possible scores to form percentage scores. This transformation ensured that IK and EK were on the same measurement scale and the descriptive statistics are shown in Table 4.41.

Table 4.41 Descriptive statistics of IK and EK percentage scores

Group	NS (n = 19)		NNS (n = 85)	
	Mean	*SD*	Mean	*SD*
IK percentage (max = 1)	.89	.04	.63	.12
EK percentage (max = 1)	.81	.08	.74	.11

It can be seen from the above table that the NS obtained a higher average score in implicit knowledge than explicit knowledge, while the NNS scored higher in explicit knowledge than implicit knowledge. This was in line with the hypotheses that the NS would have more implicit knowledge and the NNS would have more explicit knowledge. Levene's test for equality of variances was then conducted to compare the variances in implicit and explicit knowledge of the NS and the NNS.

The test results $F = 16.90$, $p < .001$ suggested that there was a statistically significant difference between the variances in implicit knowledge of the NS and the NNS groups, but a non-significant difference of the variances in explicit knowledge of the two groups ($F = 2.23$, $p = .14$). These results indicated that while the difference in the variance of explicit knowledge of the two groups was statistically nonsignificant, the variances of implicit knowledge were statistically different.

To compare the profiling of implicit and explicit knowledge of the NS and the NNS, a paired samples *t*-test was conducted for each group. For the NS, there was a statistical difference between their implicit and explicit knowledge, $t(18) = 5.20$, $p < .001$. The mean difference between the two types of knowledge was .09 with a standard deviation of .07 (95% CI: .05, .12). A significant difference was evident between the implicit and explicit knowledge of the NNS, as indicated by $t(84) = -8.47$, $p < .01$. The mean difference was -.10 with a standard deviation of .11 (95% CI: -.12, -.08). Cohen's *d* statistics were then calculated to determine the magnitude of these statistically significant differences. The results returned a $d = 1.19$ of the NS and $d = .92$ of the NNS, both of which indicated a large effect size. These results provided evidence that while the NS possessed less explicit knowledge than implicit knowledge, the NNS had more explicit knowledge than implicit knowledge.

In order to examine the extent to which the two groups differed in terms of implicit and explicit knowledge, two independent samples *t*-tests were conducted. The test statistics showed that there was a significant difference in the scores of implicit knowledge of the NNS ($M = .63$, $SD = .12$) and the NS ($M = .89$, $SD = .04$); $t(93.90) = 16.16$, $p < .01$. The mean difference was .26 (95% CI: .23, .29). There was also a significant difference in the scores of explicit knowledge of the NNS ($M = .74$, $SD = .11$) and the NS ($M = .81$, $SD = .08$), $t(102) = 2.79$, $p = .006$. The mean difference was .07 (95% CI: .02, .12). Cohen's *d* statistics were again calculated to estimate the magnitude of the differences between implicit and explicit knowledge of the NS and the NNS. The difference in implicit knowledge between the NS and the NNS had a Cohen's *d* of 2.82, which indicated a very large effect size. The

difference between the two groups in explicit knowledge was shown by a Cohen's *d* of .78, again indicating a large effect size.

4.7 Language aptitude and memory

The LLAMA test with sub-tests B, D, E, and F was used as a measure of language aptitude. Short-term memory (measured by the NonWord) and working memory (measured by the SSpan) were also considered as relevant constructs of language aptitude. The data of these tests were analysed together. Only the NNS participants (n = 86) completed these tests, among which 1 participant was identified as an outlier in the NonWord test, so the data of 85 NNS were retained for this test.

4.7.1 Correlation analysis

It has been established in Chapter 3 that all the measures of language aptitude and working memory had a satisfactory level of internal consistency (see Section 3.5). A correlation analysis was conducted to examine the relationship between these measures and the results are shown in Table 4.42.

Table 4.42 Correlation matrix for the LLAMA test and the memory tests (n = 86)

	B	D	E	F	NonWord	SSpan
B	__	.32 **	.41**	.49 **	.11	.27 *
D		__	.24*	.35 **	.08	.22 *
E			__	.24*	.07	.29 **
F				__	.04	.22 *
NonWord					__	.21
SSpan						__

** $p < .01$ * $p < .05$

It can be seen from the above table that all four sub-tests of the LLAMA test had statistically significant correlations with each other, ranging from r = .24 to r = .49. LLAMA B had stronger correlations with LLAMA E and F, which correspond to a medium (r = .41) and a large effect size (r = .49) respectively (Plonsky & Oswald, 2014). The correlations between sub-tests B and D, D and F were slightly lower (r = .32 and .35 respectively), corresponding to small to medium effect sizes (Plonsky & Oswald, 2014). Given the relatively small number of participants involved (n = 86), these correlations were considered significant. In contrast, although the correlations between sub-tests D and E, E and F were statistically significant ($p < .05$), the associated effect sizes were very small ($r < .25$).

The two tests of memory had lower correlations with the four language aptitude sub-tests in comparison to the correlations among the sub-tests of the LLAMA test. The NonWord test had no statistically significant correlations with all the other tests, as indicated by the very small r statistics. The SSpan test correlated significantly with the four LLAMA sub-tests with correlations ranging between .22 and .29. The correlation between the two memory tests had a very small effect (r = .21) that was statistically non-significant. The failure to find a significant correlation between the two memory tests was indicative of the distinct constructs that these two measured.

4.7.2 Confirmatory factor analysis

The correlation analysis results indicated that the memory tests and the LLAMA test measured different constructs. To explore the relationship between working memory and language aptitude further, a structural model that consists of two hypothesised factors (i.e. aptitude and memory) was constructed and then examined (Figure 4.8).

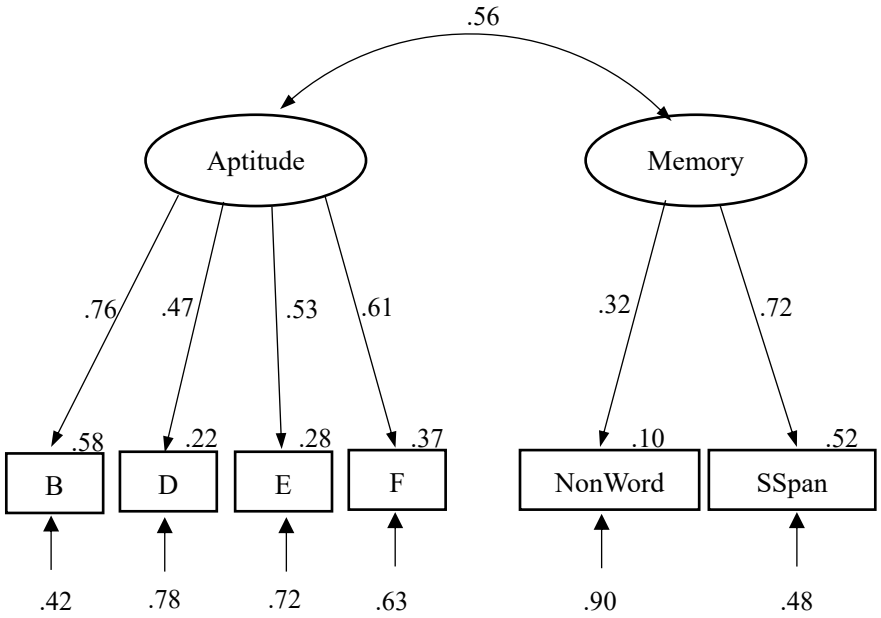

Figure 4.8 Confirmatory factor analysis model of language aptitude and memory

The KMO test was conducted to measure the sampling adequacy of this confirmatory factor analysis. The test statistics of .69 indicated adequate sampling (Kaiser, 1974). The Bartlett's test of sphericity $\chi^2(15) = 60.76$, $p < .01$, indicated that the correlations between tests were sufficient. The Mahalanobis distance of the aptitude and memory measures was calculated, followed by a calculation of its probability level using the Chi square distribution to detect multivariate outliers. According to Tabachnick and Fidell (2013), cases with Mahalanobis distance greater than the χ^2 critical value at the probability of .001 are likely to be multivariate outliers. The critical value of χ^2 at $df = 6$ and $p = .001$ is 22.458. The calculated Mahalanobis distance values ranged from .41 to 14.57, and the corresponding p value ranged from .0012 to .976. The p values were larger than the critical value of .001 and therefore it was concluded that no multivariate outliers were found in this data.

The multivariate skewness and kurtosis of the language aptitude and memory measures was checked by conducting Mardia's multivariate normality test (Mardia, 1970, 1974) using R (Korkmaz et al., 2014; R Core Team, 2017). The test results returned a Mardia's skewness of 4.58, $p = .21$ and a kurtosis of 44.91, $p = .15$,

indicating a multivariate normal distribution. The multivariate Shapiro-Wilk normality test (Royston, 1983) was also conducted to examine data normality using R (Jarek, 2012; R Core Team, 2017), and the test result $W = .97$, $p = .09$ again indicated multivariate normal distribution. Based on these results, the maximum likelihood (ML) estimation method was used in the confirmatory factor analysis.

The analysis of the hypothesised two-factor model obtained satisfactory model fit, $\chi^2(8) = 5.61$, $p = .70$. The results of other goodness-of-fit indices are presented in Table 4.43. The standardised residual covariance of this model ranged between -.94 to .54, indicating that the model was properly specified.

Table 4.43 Goodness-of-fit indices of the aptitude-memory model

GoF type	Absolute	Parsimony	Comparative	
	SRMR	RMSEA (90% CI)	CFI	TLI
Statistics	0.040	0.00 (0.00, 0.09)	1.00	1.08

The results of this CFA largely reflected the results of the correlation analysis. The four sub-tests of the LLAMA tests met the minimal level of interpretability with factor loadings ranging from .47 to .76 (Hair et al., 2014), indicating that the LLAMA test was internally consistent. However, considering the small sample size of 86, factor loadings that exceed .60 were considered to have statistical significance (Hair et al., 2014). Among the four sub-tests, LLAMA B had the highest loading of .76, which explained 58% of its total variance. LLAMA F had a loading of .61, which explained 37% of the total variance. LLAMA D and E obtained factor loadings of .47 and .53 respectively, which only explained less than 30% of their total variances. The Wald z values associated with these factor loadings ranged between 3.37 and 4.22, $p < .05$; indicating that these factor loadings were significant.

The factor loadings of the NonWord test and the SSpan test were .32 and .72 respectively, suggesting a relatively weak relationship between these two tests and the hypothesised construct memory. The SSpan test received a z statistic of 1.16 with

p = .16, indicating that this test was statistically non-significant to the overall model. The covariance between the two factors language aptitude and memory was .56, with a z value of 1.37, p = .17, again suggesting a non-significant correlation.

To sum up, the CFA results suggested that the LLAMA test and the two memory tests measured different constructs. However, although the model fit indices indicated a good overall model fit, the non-significant and the relatively low factor loadings (i.e. those below .60) in the model suggested that the strength of correlations between these measures and the hypothesised constructs was limited.

4.8 The relationship between aptitude, working memory and L2 knowledge

It was assumed that language aptitude and working memory have an impact on L2 acquisition and therefore might affect implicit and explicit L2 knowledge as well. In this section, the extent to which language aptitude and working memory influenced the implicit and explicit knowledge of the NNS was examined.

4.8.1 Correlation analysis

A correlation analysis was conducted to investigate the strength of the relationships among the aptitude test, the memory test, and the six language tests. Table 4.44 presents the Pearson correlation matrix of these tests.

Table 4.44 Correlation matrix of the LLAMA, the memory test and the language tests

	B	D	E	F	NonWord	SSpan
EI	.41 **	.17	.27 *	.49 **	.16	-.08
TGJT	.41 **	.35 **	.27*	.48 **	.17	-.06
WordM	-.04	-.21 *	-.01	-.01	-.05	-.02
UGJT	.22 *	.07	.26 *	.30 **	.00	-.02
MKT	.04	.04	.32 **	.05	.06	.21
Cloze	.19	.12	.27 *	.28 **	.10	.06

** $p < .01$ * $p < .05$

Several statistically significant correlations were observed between the four sub-tests of the LLAMA test and the language tests. LLAMA F, a test that examined the grammatical inferencing abilities of the participants, was found to correlate significantly with the EI, the TGJT, the UGJT, and the cloze test. Among these, the higher correlations were with the hypothesised tests of implicit knowledge *(r* = .49 with the EI and *r* = .48 with the TGJT), indicating medium to large effect sizes. The slightly lower yet still statistically significant correlations between LLAMA F and the UGJT (*r* = .30), and between LLAMA F and the cloze test (*r* = .28) indicated small to medium effect sizes. These results suggested a stronger relationship between LLAMA F and implicit L2 knowledge than explicit L2 knowledge.

Similarly, statistically significant correlations were found between LLAMA E, a measure of sound-symbol correspondence ability, and all the language tests except for the WordM test. However, the strengths of these correlations were all moderate with correlation coefficients below .40. The correlations between LLAMA B that measured the ability of vocabulary learning and two of the hypothesised tests of implicit knowledge were also moderate, judging from the coefficients of .41 with the EI and the TGJT. LLAMA D that measured the ability of sound recognition correlated significantly with the TGJT (*r* = .35) and the WordM (*r* = -.21).

These results indicated that while language aptitude had a correlation with L2 knowledge, working memory ability seemed irrelevant to L2 knowledge. However, rather than focusing on the relationships among individual aptitude tests, memory tests and language tests, the present study has a primary focus on understanding the overall influence of language aptitude and working memory on L2 knowledge. Therefore, it was decided that two weighted composite scores would be calculated to represent language aptitude and working memory respectively based on the results of the factor analysis (see Section 4.7.2) so that further investigations could be made to examine the influence of language aptitude and working memory on L2 learning outcomes.

The skewness and kurtosis statistics of these two scores ranged between -.26

and .24; which indicated normal data distribution. After the removal of outliers, the Kolmogorov-Smirnov test of normality statistics suggested a normal distribution of the language aptitude scores ($D(84) = .07$, $p = .20$), but a non-normal distribution of the working memory scores ($D(84) = .13$, $p = .003$). Table 4.45 summarises the descriptive statistics of these two variables.

Table 4.45
Descriptive statistics of language aptitude and working memory factor scores

Group	NNS (n = 84)	
	Mean	*SD*
Language aptitude (max = 59.25)	26.53	9.05
Working memory (max = .52)	.28	.05

Using these factor scores and the implicit and explicit knowledge factor scores (see Section 4.6), a correlation analysis was conducted to examine the relationships among them. The results are shown in Table 4.46.

Table 4.46 Correlation matrix of aptitude, memory, IK and EK

	IK	EK	Aptitude	Memory
IK	__	.55 **	.55 **	.11
EK		__	.34 **	.03
Aptitude			__	.34 **
Memory				__

** $p < .01$ * $p < .05$

The results of this correlation analysis were in line with the previous one. Language aptitude had a stronger correlation with implicit knowledge ($r = .55$) although it was also correlated significantly with explicit knowledge ($r = .34$). In contrast, there was no evidence that working memory was correlated with implicit or explicit knowledge. Language aptitude had a moderate but statistically significant correlation with working memory ($r = .34$).

To sum up, the results of these two correlation analyses indicated that language aptitude was pertinent to both implicit and explicit knowledge. Working memory, in comparison, was not connected with either type of knowledge.

4.8.2 Multiple regression analysis

Based on the results of the correlation analyses, the extent to which language aptitude influences L2 knowledge was examined further using multiple regression analyses. Using the weighted sum scores of language aptitude, working memory and implicit and explicit knowledge, two multiple regression analyses were conducted to examine if implicit and explicit knowledge can be predicted by language aptitude and working memory using the car package under R (Fox & Weisberg, 2011).

To test if the assumptions of multiple linear regressions were met, a global test that examined the multivariate skewness and kurtosis, linearity, homoscedasticity of the residuals of the regression equations was implemented using the gvlma package in R (Pena & Slate, 2014). It was found that the assumptions were satisfied for multiple regression analysis.

The analyses results indicated a statistically significant regression equation for both implicit knowledge ($F(2, 81) = 17.36$, $p < .001$ with a R^2 of .30) and explicit knowledge ($F(2, 81) = 6.25$, $p = .003$ with a R^2 of .13). Language aptitude was found to be a significant predictor of both implicit and explicit knowledge, while working memory was a non-significant predictor. Tables 4.47 and 4.48 summarise the coefficient estimates of these two models.

Table 4.47 The regression coefficients of aptitude and memory on IK

	B	*SE*	β	t	*p*	Partial R^2
(Intercept)	22.30	2.96		7.54	.000 ***	.41
Aptitude	.35	.06	.57	5.78	.000 ***	.29
Memory	-9.16	10.77	-.08	-.85	.40	.01

*** $p < .001$

Table 4.48 The regression coefficients of aptitude and memory on EK

	B	*SE*	β	t	*p*	Partial R^2
(Intercept)	22.74	2.13		10.67	.000 ***	.58
Aptitude	.16	.04	.39	3.53	.001**	.13
Memory	-7.07	7.76	-.10	-.91	.36	.01

*** $p < .001$ ** $p < .05$

The positive coefficient estimate of language aptitude in Table 4.47 indicated a positive linear relationship between aptitude scores and implicit L2 knowledge. The increase of .35 in aptitude score would bring about an increase of 1 point in implicit knowledge. Similarly, the increase of .16 in aptitude score would bring about an increase of 1 point in explicit knowledge. The much larger beta estimates of language aptitude (β = .57 and .39) than those of working memory (β = -.08 and -.10), suggested that language aptitude was a significant predictor of both implicit and explicit knowledge. Figures 4.9 and 4.10 depict these relationships.

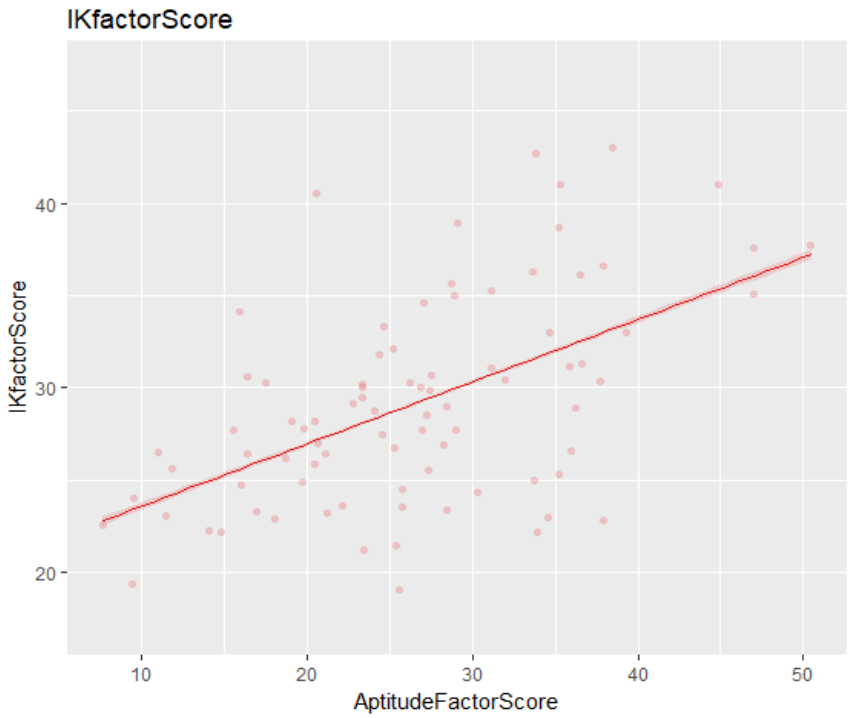

Figure 4.9 The influence of language aptitude on implicit L2 knowledge

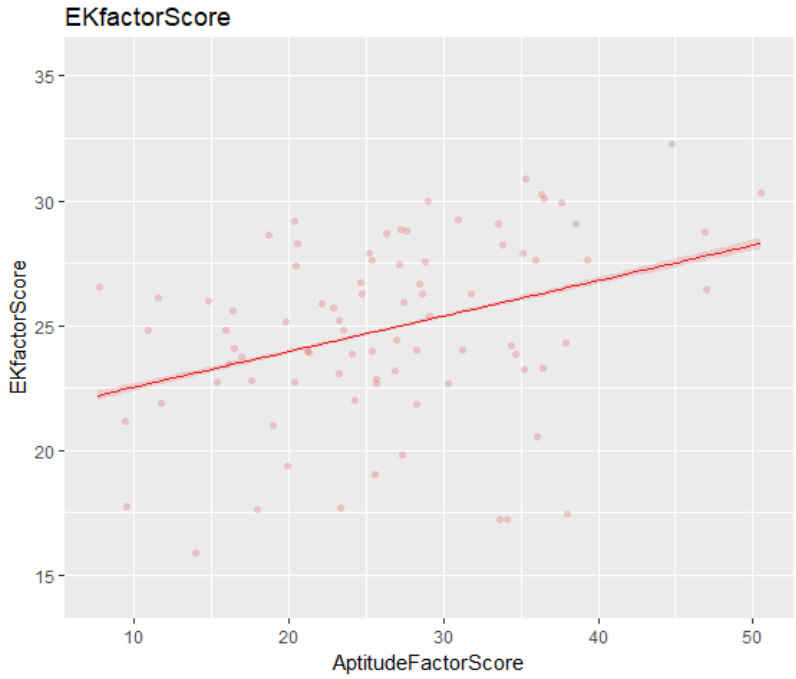

Figure 4.10 The influence of language aptitude on explicit L2 knowledge

4.9 Age and L2 knowledge

In the present study, the data of three variables related to learning age were collected to answer the research question of how age influences learning outcomes. These three variables were: 1) AO (age of onset) - the age when the learning of English started, 2) AoA (age of arrival) - the age of immigration to New Zealand, and 3) AoT (age of testing) - the age of participation in the present study. Another closely relevant variable, LoR (length of residence) in the L2 country was also analysed together with the age variables.

4.9.1 Correlation analysis

A correlation analysis was conducted to examine the relationship between the three age variables, LoR and L2 knowledge, and the results are shown in Table 4.49.

Table 4.49 Correlation matrix of the age variables, LoR and L2 knowledge

	AoA	AO	AoT	LoR	IK	EK
AoA	__	.70 **	.92**	.02	-.53 **	.09
AO		__	.69**	-.10	-.48 **	.05
AoT			__	.36 **	-.50 **	.12
LoR					.08	.16
IK					__	.55**
EK						__

** $p < .01$

It can be seen from these results that the age variables had significant correlations that ranged between -.48 and -.53 with implicit L2 knowledge, but not with explicit knowledge. The correlations among the age variables were significant as well, as indicated by coefficients ranged between .69 and .92. However, the very high correlation between AoA and AoT (r = .92, p < .01) also indicated multicollinearity, that is, the inter-associations between the two were too high (Tabachnick & Fidell, 2013). This means that if AoA and AoT were both specified as predictors in a regression analysis, the statistical power of the regression analysis would be undermined. Therefore, it is generally advised that only one of the highly correlated variables would be included in regression analysis (Field, 2009; Tabachnick & Fidell, 2013).

4.9.2 Multiple regression analysis

A multiple linear regression model that predicted implicit L2 knowledge based on AoA, AO and LoR was constructed using the car package (Fox & Weisberg, 2011) under the R environment (The R Core Team, 2017). Because AoT was too highly correlated with AoA, it was left out (Field, 2009; Tabachnick & Fidell, 2013). By performing a global test of regression assumptions using the gvlma package in R

(Pena & Slate, 2014), it was concluded that the assumptions of multiple regression were met.

A significant regression model AgeIK was found, $R^2 = 0.31$, $F(3, 82) = 12.54$, $p < .001$. The standardized residuals of this model ranged between -2.52 and 2.22, which were below the critical values of ± 3 (Tabachnick & Fidell, 2013).

Table 4.50 summarises the coefficient estimates of this model. As can be seen, AoA was statistically significant (β = -.36, p = .006), whereas AO was not statistically significant (β = -.24, p = .066). This indicated that AoA had a stronger and substantial influence on implicit knowledge, while AO did not have such an influence. Length of residence also did not influence implicit L2 knowledge significantly (β = .11, p = .236).

Table 4.50 Coefficient estimates of model AgeIK

	B	*SE*	β	t	*p*	Partial R^2
(Intercept)	35.67	2.42		14.72	.000 ***	.73
AoA	-.20	.07	-.36	-2.84	.006 **	.09
AO	-.44	.24	-.24	-1.86	.066	.04
LoR	.14	.12	.11	1.20	.236	.02

*** $p < .001$ ** $p < .05$

Figure 4.11 demonstrates the relationship between AoA and implicit knowledge. A negative linear relationship between AoA and implicit knowledge was evident. As learners' age of immersion in the L2 contexts increased, there was a linear decline in their scores of implicit L2 knowledge.

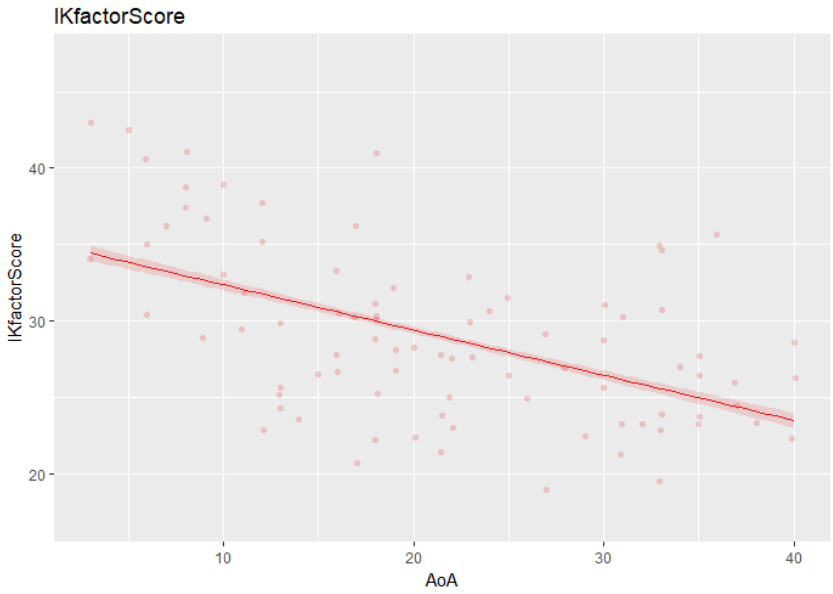

Figure 4.11 Predicted probabilities of IK by AoA

Similarly, AoA, AO, and LoR were examined as predictors of explicit L2 knowledge in the regression model AgeEK. As anticipated, a non-significant regression equation was yielded, $F(3, 81) = 1.07$, $p = .37$. Only 4% of the variance in explicit knowledge was explained by these variables ($R^2 = .04$). The two age variables were not significant predictors of explicit knowledge, nor was length of residence. The coefficient estimates of this analysis are presented in Table 4.51.

Table 4.51 Coefficient estimates of model AgeEK

	B	*SE*	β	t	*p*	Partial R^2
(Intercept)	22.53	1.88		11.96	.000 ***	.64
AoA	.05	.06	.14	.89	.377	.01
AO	-.08	.19	-.06	-.42	.677	.00
LoR	.14	.09	.17	1.55	.125	.03

*** $p < .001$

Another possibility of the influence of age was a non-linear influence. In order to test this hypothesis, AoA was recoded into different groups. In line with the literature (DeKeyser, 2013), AoAs below 13 were coded as young; AoAs between 14

and 20 were coded as teen; 21 to 30 were coded as adult; and beyond 30 were coded as late. This categorization yielded four more or less equal groups, n = 23, 20, 21, 22 respectively. The descriptive statistics of the IK and EK according to age groups are shown in Table 4.52. Then a regression model AgebyGIK was once again built to predict implicit knowledge based on AoA groups, and the model estimates were presented in Table 4.53.

Table 4.52 Descriptive statistics of IK and EK according to age groups

AoA group	IK		EK	
	Mean	*SD*	Mean	*SD*
Young (AoA ≤13)	33.83	5.97	24.52	4.14
Teen (13<AoA≤20)	28.64	4.78	24.72	4.06
Adult (20<AoA≤30)	26.72	3.66	24.50	3.25
Late (AoA>30)	26.29	4.52	25.90	2.99

Table 4.53 Coefficient estimates of model AgebyGIK

	B	*SE*	β	t	*p*
(Intercept)	33.83	1.01		33.49	.00 ***
AoA-Teen	-5.19	1.46	-.40	-3.55	.00 ***
AoA-Adult	-7.11	1.48	-.54	-4.80	.00 ***
AoA-Late	-7.54	1.45	-.59	-5.22	.00 ***

*** $p < .001$

This model was significant with an $R^2 = .29$, $F(3, 82) = 11.42$, $p < .001$. The results of this model largely mimicked the results of model AgeIK and indicated a linear declining trend of implicit knowledge as AoA increased. This trend is portrayed in Figure 4.12.

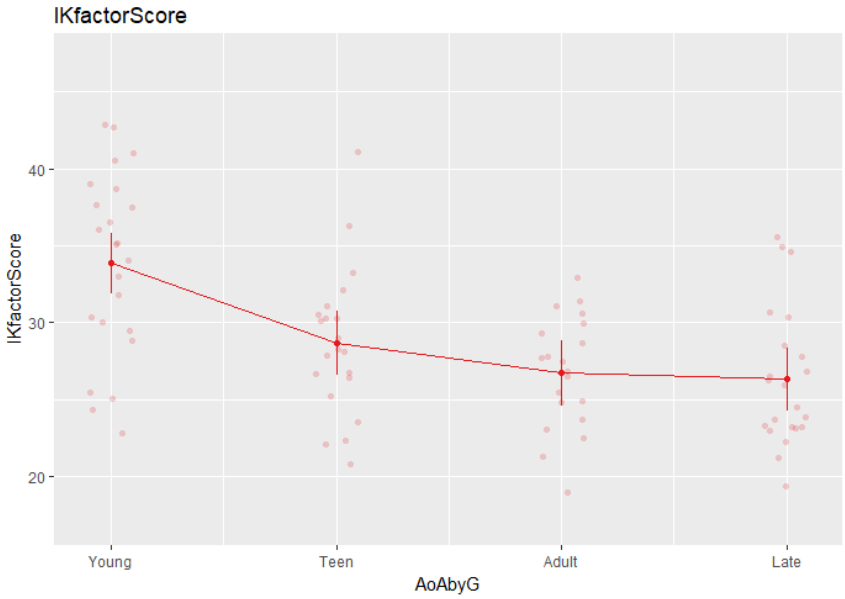

Figure 4.12 Predicted probabilities of IK by AoA groups

4.10 The influence of age and aptitude on L2 knowledge

4.10.1 The joint influence of age and aptitude

So far in the data analysis, the influence of age and aptitude on L2 knowledge has been examined separately. The next question of interest was to investigate how these two factors jointly influenced L2 knowledge.

A multivariate multiple regression model AAM1 was conducted to examine the influence of AoA and language aptitude on implicit and explicit L2 knowledge. This analysis is an extension of multiple regressions by allowing for more than one dependent variable. The analysis results showed that age and aptitude were both significant predictors of L2 knowledge, Pillai's trace = .45, F (2, 81) = 33.13, p < .001 for age; and Pillai's trace = .20, F (2, 81) = 10.40, p < .001 for language aptitude. The partial R^2 for age and aptitude were .44 and .20 respectively.

The univariate regression analysis results indicated a significant regression equation for implicit L2 knowledge, F (2, 82) = 29.56, p< .001 with an R^2 of .42. Linear regression assumptions were checked by performing a global test using the

gvlma package in R (Pena & Slate, 2014), and the results suggested that all the assumptions were met. Table 4.54 below presents the coefficient estimates of this model, and Figure 4.13 visually illustrates the influence of age and language aptitude.

Table 4.54 Coefficient estimates of model AAM1 on IK

	B	*SE*	β	t	*p*	Partial R^2
(Intercept)	27.22	2.27		12.01	.000 ***	.64
AoA	-.21	.05	-.38	-4.09	.000 ***	.17
Aptitude	.24	.06	.39	4.20	.000 ***	.18

*** $p < .001$

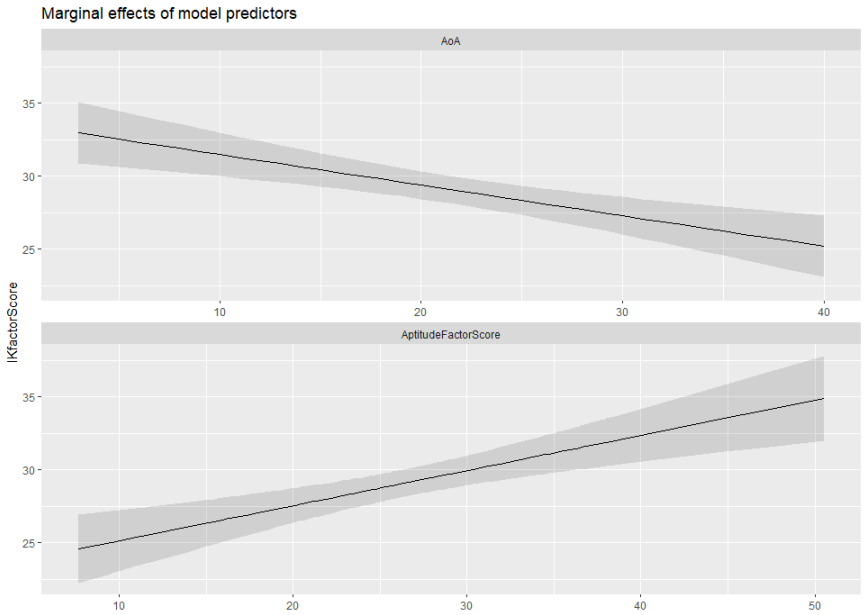

Figure 4.13 The influence of age and aptitude on IK

In the analysis of the influence of age and aptitude on explicit knowledge, the assumptions of normal distribution of residuals and homoscedasticity were violated. This was caused by the negatively skewed data of explicit L2 knowledge; therefore a square root transformation was performed as a remedy. Then the regression analysis was run again and this time the model passed all the tests of regression assumptions.

The analysis results again returned a significant regression equation, F (2, 82) = 10.57, p < .001 with an R^2 of .20. AoA was found to be a significant predictor of explicit knowledge (β = -.28, p = .011), so was language aptitude (β = –.49, p < .001). Table 4.55 below presents the coefficient estimates of this model, and Figure 4.14 visually illustrates the influence of age and language aptitude on explicit knowledge.

Table 4.55 Coefficient estimates model AAM1 on EK

	B	*SE*	β	t	*p*	Partial R^2
(Intercept)	4.29	.27		15.79	.000 ***	.75
AoA	-.02	.01	-.28	-2.60	.011 *	.08
Aptitude	-.03	.01	-.49	-4.53	.000 ***	.20

*** p < .001

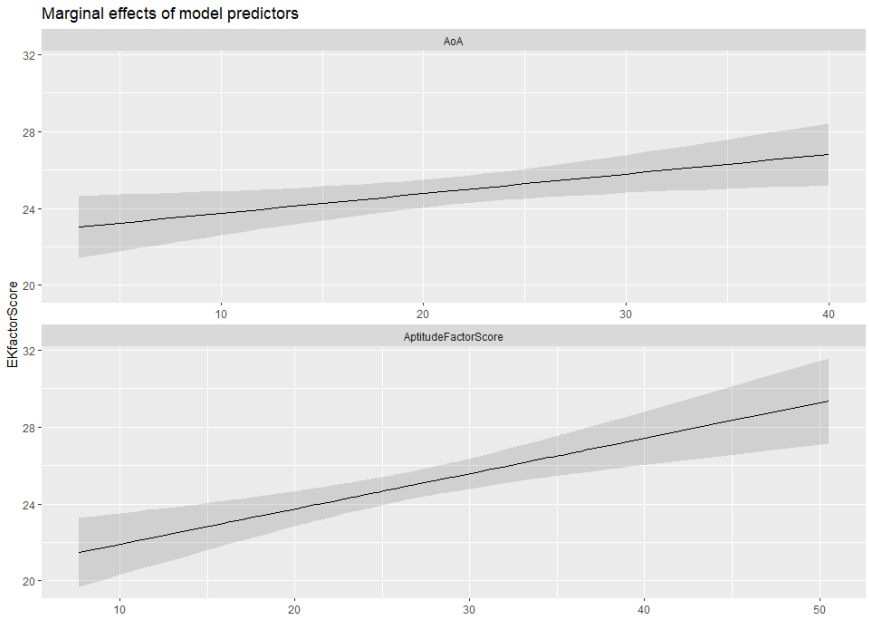

Figure 4.14 The influence of age and aptitude on EK

Taken together, the results of the above two analyses demonstrated that although age and language aptitude were both significant predictors of L2 knowledge, they were of different degrees of relevance to implicit and explicit knowledge. These two variables contributed similarly in predicting implicit L2

knowledge, as indicated by the similar standardised coefficient estimates (β = -.38 and .39) and the partial correlation R^2 statistics (R^2 = .17 and .18). In comparison, language aptitude had a stronger predictive power for explicit knowledge, as indicated by the larger standardised coefficient estimates (β = -.49) and the larger partial correlation estimates (R^2 = .20). These results were partly in line with the previous literature, showing that language aptitude was a relevant construct to explicit learning and explicit knowledge. In addition, these results suggested that language aptitude also played a role in the development of implicit L2 knowledge. Age on the other hand, had a closer link to implicit knowledge than explicit knowledge.

4.10.2 The interaction between aptitude and age

The next research question of interest was to investigate whether there was an interaction between age and language aptitude. However, in multiple regression analysis, the interaction effects between two continuous variables become difficult to interpret (Jeon, 2015). Therefore, to facilitate interpretability, AoA was categorised into four groups (see Section 4.9.2), as was language aptitude. The numerical scores of the LLAMA test were recoded into four grade scores (i.e. below average, average, good, outstanding) according to the LLAMA manual (Meara, 2005).

A multivariate analysis of variances (MANOVA) model AAM2 was constructed to predict implicit and explicit L2 knowledge based on the main effects of AoA and aptitude. The young group was selected as the reference level for AoA, and the below average group for aptitude.

Since the data was unbalanced (i.e. the number of observations for each level of a variable was not equal), Type II sum of squares MANOVA test results were interpreted (Fox & Weisberg, 2011; Huberty & Olejnik, 2006). It was found that the age factor was significant, Pillai's trace = .50, F (6, 156) = 8.72, p < .001; so was language aptitude, Pillai's trace = .21, F (6, 156) = 3.09, p = .007. The partial η^2

were .25 and .11 for age and aptitude respectively.

The univariate analysis of the influence of AoA and aptitude on implicit knowledge showed a significant regression equation, F (6, 78) = 7.96, p < .001. Table 4.56 presents the results of this analysis. It can be seen that the age factor had a significant influence on implicit knowledge. Compared to the young participants who arrived in New Zealand before the age of 13, the other three groups obtained significantly lower scores in implicit knowledge (β = -.26, p = .021; β = -.39, p = .001; and β = -.44, p < .001 for the teen, the adult and the late group, respectively). In other words, AoA had a negative influence on implicit knowledge, meaning that the later one arrived in the L2 environment, the less implicit knowledge developed.

The influence of language aptitude on implicit knowledge was also significant. In comparison to the below average group, the average group did not show much difference (β = .30, p = .054), but the other two groups demonstrated significant increases (β = .39, p = .014 and β = .34, p = .005 for the good and the outstanding group, respectively). These results indicated that an above average level of aptitude would result in better achievement in implicit knowledge. The partial η^2 for age and language aptitude and were .17 and .11 respectively, and the overall R^2 of the model was .38.

Table 4.56 Univariate coefficient estimates of model AAM2 on IK

	B	*SE*	β	t	*p*
(Intercept)	28.78	2.07		13.90	.000 ***
AoA-Teen	-3.49	1.49	-.26	-2.35	.021 *
AoA-Adult	-5.14	1.56	-.39	-3.30	.001 **
AoA-Late	-5.61	1.53	-.44	-3.67	.000 ***
Aptitude-Average	3.44	1.76	.30	1.95	.054
Aptitude-Good	5.08	2.03	.39	2.51	.014 *
Aptitude-Outstanding	9.02	3.09	.34	2.92	.005 **

*** p < .001 ** p < .01 * p < .05

The univariate analysis of the influence of AoA and aptitude on explicit knowledge also returned significant results, $F(6, 78) = 3.80$, $p = .002$. However, the pattern of the influence of these two factors was different, as was shown in Table 4.57. Again, AoA was found to be a significant factor, but it had a positive influence on explicit knowledge. Compared to the young group, the teen and the adult group did not show significant increases in the scores of explicit knowledge ($\beta = .19$, $p = .127$, and $\beta = .23$, $p = .082$, respectively), while the late group showed a significant difference ($\beta = .41$, $p = .003$).

Language aptitude, again, was found to have had a significant influence. With higher levels of language aptitude, the scores of explicit knowledge differed substantially across groups ($\beta = .43$, $p = .015$; $\beta = .53$, $p = .003$; and $\beta = .57$, $p = .001$ for the average, the good, and the outstanding aptitude group respectively). These results indicated that differences in language aptitude caused differences in the scores of explicit knowledge, suggesting a significant role of language aptitude. In comparison, the role of age was less evident. The partial η^2 of age and language aptitude were.11 and .21 respectively; the overall R^2 of explicit knowledge was .23.

Table 4.57 Univariate coefficient estimates of model AAM2 on EK

	B	*SE*	β	t	*p*
(Intercept)	19.68	1.49		13.18	.000 ***
AoA-Teen	1.65	1.07	.19	1.54	.127
AoA-Adult	1.98	1.13	.23	1.76	.082
AoA-Late	3.36	1.10	.41	3.04	.003 **
Aptitude-Average	3.18	1.27	.43	2.50	.015 *
Aptitude-Good	4.48	1.46	.53	3.07	.003 **
Aptitude-Outstanding	9.77	2.23	.57	4.37	.001 ***

*** $p < .001$ ** $p < .01$ * $p < .05$

The following HE plot (Figure 4.15) demonstrates the joint influence of age and aptitude on implicit and explicit L2 knowledge (Fox & Weisberg, 2011; Friendly, 2007). The plot was drawn using the heplots package (Fox, Friendly, & Monette, 2017) under the R environment (The R Core Team, 2017). In MANOVA models, an H matrix and an E matrix were produced as a result of the calculation of univariate sums of squares and the error variances, respectively. The covariance of variables within the model was represented by ellipses, which demonstrate the relationship between the H matrix and the E matrix. In model AAM2, the E matrix was represented by the red dashed ellipse, and the magenta and blue ellipses represented the effects of age and aptitude respectively. The effects of the two factors in the model were rescaled in relation to Roy's largest root statistics (a general MANOVA test statistic), so that the significance of them could be shown. When the ellipses of factors protrude the E ellipse, the significance of these factors is implied. Otherwise, it indicates a non-significant effect (Friendly, 2010). In model AAM2, both age and aptitude were significant factors, as were shown in Figure 4.15 by the two corresponding ellipses protruding the error ellipse.

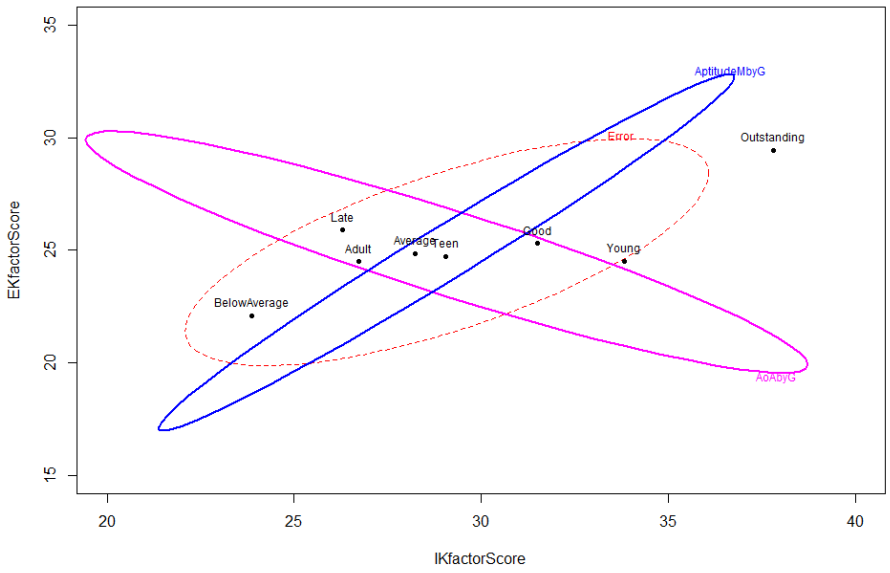

Figure 4.15 HE plot of model AAM2

It can be seen that overall AoA had a stronger effect (i.e. the large magenta ellipse) than language aptitude (i.e. the small blue ellipse). The shape of the ellipses also indicated the strengths of the influence of these factors. The magenta ellipse spanned across more units along the x axis than the y axis, indicating a stronger influence of AoA on implicit knowledge. In comparison, the diagonally placed blue ellipse of language aptitude demonstrated similar influence to implicit and explicit L2 knowledge.

Differences among age groups and aptitude groups also appeared in Figure 4.15. The group means were represented by labelled dots and it can be seen that there were substantial differences among age groups. For example, while the young group had an estimated mean score of 34 in implicit knowledge, and 25 in explicit knowledge; the late group had a mean score of 26 in implicit knowledge and explicit knowledge. The distance between the two dots was a visual representation of the vast difference between these two groups. Likewise, large differences can be seen among other age groups and aptitude groups.

Apart from the above examined joint effects of age of immersion and language aptitude, the interaction effect of the two might play a role in language learning too. In order to investigate the interaction effect, a new MANOVA model AAM3 was constructed by including an interaction independent variable AoA*Aptitude on top of model AAM2. Then an ANOVA *F* test was conducted to compare models AAM2 and AAM3. The test results $F(10, 146) = .47$, $p = .91$, indicated that the addition of the interaction effect did not produce significant improvement to model AAM2.

The usefulness of this interaction between age and aptitude was also assessed by checking the adjusted R^2 statistics of model AAM3 in comparison to model AAM2 (Table 4.58). The adjusted R^2 is a modified version of R^2 according to the number of predictors in the model. When a newly added predictor improves the model more than would be expected by chance, the adjusted R^2 would increase (Pituch & Stevens, 2016). It can be seen that in model AAM3, the adjusted R^2 of both implicit and explicit knowledge decreased with the addition of the interaction

factor. This means that the addition of this interaction did not lead to an improvement of the model.

Table 4.58 Comparison of models AAM2 and AAM3

Model	R^2		Adjusted R^2	
	IK	EK	IK	EK
AAM2	0.38	0.23	0.33	0.17
AAM3	0.40	0.27	0.31	0.16

Some possible combinations of the age group and aptitude group did not exist in the data, these were: the late group with a good level of aptitude, and the teen, the adult, and the late group with an outstanding level of aptitude. Therefore, the partial η^2 of individual variables were not successfully calculated. The univariate coefficient estimates of model AAM3 on implicit and explicit knowledge are shown in Tables 4.59 and 4.60.

Table 4.59 Univariate coefficient estimates of model AAM3 on IK

	B	*SE*	β	t	*p*
(Intercept)	22.18	2.70		8.22	.000 ***
AoA-Young	6.96	3.65	.55	1.91	.061
AoA-Teen	2.18	4.27	.17	0.51	.611
AoA-Adult	3.07	3.81	.23	0.80	.424
Aptitude-Average	4.80	2.93	.42	1.64	.104
Aptitude-Good	4.44	4.27	.34	1.04	.301
Aptitude-Outstanding	8.66	5.11	.33	1.70	.094
Young:Average	-1.47	4.10	-.08	-.36	.721
Teen:Average	-.13	4.65	-.01	-.03	.98
Adult:Average	-3.76	4.16	-.26	-.90	.37
Teen:Good	1.41	5.64	.07	.25	.80
Adult:Good	2.50	6.03	.07	.42	.68

*** $p < .001$

Table 4.60 Univariate coefficient estimates of model AAM3 on EK

	B	*SE*	β	t	*p*
(Intercept)	21.65	1.92		11.26	.000 ***
AoA-Young	-3.23	2.60	-.40	-1.24	.219
AoA-Teen	-1.27	3.04	-.15	-.42	.677
AoA-Adult	2.03	2.72	.24	0.75	.458
Aptitude-Average	4.87	2.09	.66	2.33	.022 *
Aptitude-Good	5.40	3.04	.64	1.77	.080
Aptitude-Outstanding	11.02	3.64	.65	3.03	.003 **
Young:Average	-.11	2.92	-.01	-.04	.969
Teen:Average	-.40	3.32	-.04	-.12	.904
Adult:Average	-4.47	2.97	-.47	-1.51	.136
Teen:Good	-.14	4.02	-.01	-.03	.972
Adult:Good	-.22	4.30	-.01	-.05	.959

*** $p < .001$ ** $p < .01$ * $p < .05$

It can be seen from the above tables that the all the interactions between age and aptitude were non-significant. Moreover, the addition of the interaction also seemed to mask the independent influence of age and language aptitude. Therefore, based on these results and the model comparison results, it was concluded that the effect of the interaction between AoA and aptitude was trivial. However, AoA and aptitude had a significant impact on L2 knowledge as independent factors (see Section 4.10.1). AoA influenced implicit and explicit knowledge differently. When AoA increased, the amount of implicit knowledge decreased, accompanied by an increase in the amount of explicit knowledge. In contrast, language aptitude had a positive influence on both implicit and explicit knowledge. The higher level of aptitude, the more implicit and explicit knowledge were developed.

4.11 Some near-native NNS

Within the 86 NNS participants, a small number of participants (n = 9) were able to score within the NS range in implicit knowledge (IK of the NS: max = 43.79, min = 37.39). The scores of their implicit knowledge are listed together with their AoA, language aptitude, LoR and years of education that they received in New Zealand in Table 4.61.

It can be seen from this table that out of the 9 near-native NNS participants, 8 had an AoA before puberty (i.e. AoA ≤ 13). The only exception was participant 4, who had an AoA of 18 years old. With respect to language aptitude, seven participants had an above average level of aptitude. Within these 7 participants, 3 of them had an excellent level of language aptitude.

Table 4.61 Information about the near-native NNS (n = 9)

Participant No.	IK (max=45.72)	AoA	Aptitude	Aptitude -Group	LoR	YrsinNZ Edu
1	42.95	3	38.54	Good	20	11
85	42.58	5	33.84	Good	15	10
86	41.01	8	44.81	Excellent	12	8
4	41.01	18	35.28	Good	14	8
74	40.54	6	20.48	Average	16	9
46	38.96	10	29.05	Average	16	12
62	38.71	8	35.14	Good	8	7
76	37.65	12	50.53	Excellent	15	11
67	37.50	8	46.89	Excellent	27	12

Since these near-native participants demonstrated both an early age of immersion and a higher level of language aptitude, the researcher decided to take a closer look at the participants who arrived in New Zealand before the age of 13 but failed to be near-native in implicit knowledge in order to further understand the roles

of AoA and aptitude. There were altogether 23 participants (i.e. the young group, see Section 4.9.2), who were immersed in the English environment before (including) 13 years old, and 15 of them did not score within the NS range in implicit knowledge. Table 4.62 presents the IK scores, AoA, language aptitude scores, LoR and years of education received in New Zealand of these participants.

Table 4.62 Information about the young NNS who were not near-native (n = 15)

Participant No.	IK (max=45.72)	AoA	Aptitude	Aptitude -Group	LoR	YrsinNZ Edu
81	36.59	9	37.99	Good	15	8
69	36.11	7	36.49	Good	14	16
58	35.24	12	34.70	Good	12	12
71	35.02	6	46.93	Excellent	26	10
73	34.04	3	16.00	Average	20	15
75	33.01	10	31.25	Average	13	7
65	31.81	11	24.38	Average	10	6
68	30.31	6	17.58	Average	14	9
33	29.91	12	26.95	Average	15	10
61	29.47	11	23.31	Average	10	11
64	28.88	9	28.36	Average	10	9
83	25.53	13	27.32	Average	9	10
82	25.09	13	33.75	Good	8	11
44	24.32	13	30.35	Good	18	8
80	22.86	12	38.03	Good	14	10

It can be seen from Table 4.62 that the scores of participants 81, 69, 58 and 71 in implicit knowledge were close to the lowest NS score (NS min = 37.39). These four participants all had a relatively high level of language aptitude, three being on the good level and one on the excellent level. The other eleven young participants

scored lower than 35 in implicit knowledge, with the lowest being 22.86 (participant 80). The majority (n = 8) of these eleven participants had an average level of language aptitude. However, although subjects 82, 44 and 80 had a good level of language aptitude, their implicit knowledge scores were far from the NS range.

4.12 Questionnaire data

4.12.1 Demographic information of the NNS

Demographic information about the NNS participants was gathered by using the questionnaire described in Section 3.4.6. Table 4.63 below presents a summary of this information. This includes: the beginning age of studying English in China (AO), the age of immersion in the English-speaking context in New Zealand (AoA), the age of participation in the present study (AoT) and length of residence in the L2 country (LoR). Information about the length of learning English in different levels of education in China and the frequency of English classes per week is also provided. After immigration, some NNS participants attended primary/secondary schools and universities, so information about the length of their education in New Zealand is also provided.

In the Chinese education system, students are required to complete 9 years of compulsory education, including 6 years in primary schools and 3 years in junior secondary schools (i.e. middle school). Students then continue to study in senior secondary schools (i.e. high school) for 3 years before entering universities. For some of the NNS in the present study, English was introduced in middle schools and taught as a compulsory course in secondary schools and universities. However, nowadays English is typically introduced to students early in primary schools.

Table 4.63 Demographic information of the NNS (*n* = 86)

Context	Demographics	Mean	*SD*
China	Age of onset (*n* = 86)	10.15	3.05
NZ	Age of arrival (*n* = 86)	21.40	10.10
NZ	Age of testing (*n* = 86)	36.98	11.08
NZ	Length of residence (*n* = 86)	15.32	4.41
China	Years of learning English in primary school (*n* = 75)	1.29	1.72
China	Years of learning English in middle school (*n* = 68)	2.69	.74
China	Years of learning English in high school (*n* = 58)	2.64	.72
China	Years of learning English in university (*n* = 40)	2.70	.97
China	No. of English classes per week in primary school (*n* = 75)	1.30	1.90
China	No. of English classes per week in middle school (*n* = 68)	3.09	2.36
China	No. of English classes per week in high school (*n* = 58)	3.20	2.96
China	No. of English classes per week in university (*n* = 40)	7.70	7.95
NZ	Years of education in primary school (*n* = 20)	4.8	2.75
NZ	Years of education in secondary school (*n* = 34)	4.0	1.37
NZ	Years of education in university (*n* = 77)	3.47	1.98
NZ	Years of using English in a work context (*n* = 83)	7.64	5.35

4.12.2 Learning experience of the NNS

As well as the demographic information, a variety of information regarding the learning experience of the NNS participants was collected based on the responses of the participants to a series of multiple-choice questions. These questions were designed so that inferences about the type of learning that took place could be made.

The researcher asked whether participants' learning was based on textbook or communication (see question 7, Appendix 8). It was hypothesised that the former kind of learning would suggest the implementation of more explicit learning, whereas the latter suggests more implicit learning. Other questions, including

teachers' language of instruction and the number of students in English classes were asked (see Appendix 8). As the majority of English teachers in the Chinese education system were non-native speakers (i.e. English teachers who speak Chinese as L1), the researcher also asked whether the NNS had had experience of learning English with teachers who were native speakers of English. It was hypothesised that these two kinds of teachers would adopt different approaches to teaching English, and questions 7 and 17 were asked of the NNS about teachers' instruction approaches.

In addition, an attempt was made to differentiate deductive and inductive instruction in the questionnaire. Questions 8 to 11 were designed specifically to ask about the type of instruction received by the NNS (see Appendix 8), that is, whether the teaching of English grammar, vocabulary, speaking and listening was done in a deductive or inductive way. Although both deductive and inductive instruction are considered explicit in nature (R. Ellis, 2008b), deductive instruction involves a higher degree of learners' awareness and therefore is generally believed to be more explicit than deductive instruction. Therefore, by asking this series of questions, the researcher would be able to infer the learning processes that the NNS were involved in. For example, in question 8 that asked about teachers' instruction method of grammar, the two choices given for the participants were: (A) The teacher explained the grammar rules first, and then asked me to memorize; and (B) The teacher gave sample sentences and asked me to find out the grammar rules by myself. Choice (A) was designed as a description of deductive instruction, in which learners are generally given a rule first and then asked to practise with it. Choice (B) was intended as a description of inductive instruction, in which learners are required to induce rules from examples given to them. Similarly, in questions 9, 10 and 11, choice (A) was designed as a description of deductive instruction, whereas choice (B) a description of inductive instruction.

Table 4.64 provides a summary of the relevant learning experience of the NNS in China. The column named "No." shows the number of participants who chose a particular option and the last column shows the percentage. The number of

participants who provided an answer to each question is also shown in the table.

Table 4.64 English language learning experiences of the NNS in China

Learning experiences		No.	%
No. of students in class	Less than 30	4	6.1
(*n* = 66)	30 to 40	12	18.2
	40 to 50	24	36.4
	More than 50	26	39.4
NNS teachers' language of instruction	More English	4	6.3
(*n* = 64)	More Chinese	45	70.3
	Both	15	23.4
NNS teachers' approach to teaching English	Textbook/Grammar	65	92.9
(*n* = 70)	Communication	5	7.1
NNS teachers' teaching method of grammar	Deductive instruction	67	95.7
(*n* = 70)	Inductive instruction	3	4.3
NNS teachers' teaching method of	Deductive instruction	65	92.9
Vocabulary	Inductive instruction	5	7.1
(*n* = 70)			
NNS teachers' teaching method of speaking	Deductive instruction	67	95.7
(*n* = 70)	Inductive instruction	3	4.3
NNS teachers' teaching method of listening	Deductive instruction	64	91.4
(*n* = 70)	Inductive instruction	6	8.6
Experience of learning with a NS teacher	Yes	14	20
(*n* = 70)	No	56	80
NS teachers' approach to teaching English	Grammar	1	7.1
(*n* = 14)	Communication	13	92.9
Major in English in universities	Yes	14	35
(*n* = 40)	No	26	65

The data presented in this table provides some evidence that the NNS primarily learnt English explicitly in China. Due to the large number of students in the English classes, it was probably very difficult for the teachers to adopt a teaching approach that emphasised communication. As reported by almost all NNS (92.9%), their learning of English was textbook-based with a focus on grammar. The majority of the NNS learnt English with Chinese English teachers, who generally used Mandarin

Chinese as the language of instruction and taught deductively. A small number of NNS participants (20%) had some experience of learning English with a NS English teacher, and they reported that they mainly learnt to express their ideas in English with their NS teacher.

A range of questions were designed to gather information about the learning experience and language use situation of the NNS in New Zealand. The participants were asked to report the level of difficulty they experienced in using English for the purposes of communication, reading and writing at arrival in New Zealand (questions 24 and 25, see Appendix 8). The researcher also asked the participants about their learning experience in language schools, and teachers' teaching approach in language schools. Table 4.65 below summarises this information.

Table 4.65 English language learning experience of the NNS in New Zealand

Learning experience		No.	%
Self-rated speaking/communication ability at arrival in NZ (n = 86)	No problem	9	10.5
	Sometimes difficult	28	32.6
	Difficult	49	57
Self-rated reading and writing abilities at arrival in NZ (n = 86)	No problem	26	30.2
	Sometimes difficult	28	32.6
	Difficult	32	37.2
Attending English language schools (n = 86)	Yes	26	30.2
	No	60	69.8
Approach to teaching English in language schools (n = 26)	Grammar	4	15.4
	Communication	22	84.6

It can be seen that most of the NNS encountered difficulties in speaking, reading and writing upon arrival in New Zealand. They found that speaking was more difficult than reading and writing, as 57% rated speaking being difficult for them in comparison to 37.2% who rated reading and writing as difficult. About 30% of the NNS had the experience of attending language schools, but the majority of

these participants reported that their instruction was oriented towards a focus on communicative competence, rather than an emphasis on grammatical knowledge and accuracy.

4.12.3 Language use situation of the NNS

The last question in the questionnaire was designed to ascertain the current language use situation of the NNS. A range of activities were listed as choices and the participants were asked to select all the ones that applied to them at the time of their participation in the study. The data gathered from this question is shown in Table 4.66.

Table 4.66 Language use of the NNS (n = 86)

Language use	No.	%
I go to an English-speaking church regularly.	12	14.0
I socialize with my English-speaking friends regularly.	56	65.1
I mainly socialize with other Chinese people.	64	74.4
I speak more Chinese than English with my family at home in NZ.	72	83.7
I speak more English than Chinese with my family at home in NZ.	9	10.5
I speak English with my children at home in NZ.	9	10.5
I speak Chinese with my children at home in NZ.	33	38.4
I speak English and Chinese with my children at home in NZ.	26	30.2
I like to attend local activities and talk to Kiwis in English.	43	50.0
I often watch NZ TV or listen to local radio programmes in English.	49	57.0
I often visit NZ news websites or read English newspapers (e.g. NZ herald).	48	55.8
I often watch English TV shows/movies, including those from US, UK, etc.	52	60.5

Of the total number of responses, 83.7% of NNS participants reported that they use Chinese at home with their family and 74.4% said that they mainly socialise with

other Chinese people. However, the participants also reported that they engage in activities where they use English, for example, to attend to local activities and talk to Kiwis (50%), or to watch English TV shows and movies (60.5%), listen to local radio programmes (57%), and read English newspapers (55.8%).

The results of the present study, including descriptive and inferential statistics results, and the results of the questionnaire, have been presented in this chapter. In the next chapter, these results are discussed in detail.

CHAPTER FIVE

DISCUSSION

In this chapter, the results of the present study are discussed in relation to the four research questions. However, as all the research questions are based on the premise that different types of L2 knowledge, that is, implicit and explicit knowledge, can be accessed independently, it follows that the validation of the tests of implicit and explicit knowledge are of critical importance. Therefore, some discussion regarding the validity of all the language measures is presented first. Then, the construct validity of the measures of language aptitude and working memory is discussed. Finally, the research questions are addressed.

5.1 Validation of test instruments of implicit and explicit knowledge

The validity of a test refers to "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). The validity of a certain test can be assessed in multiple ways, including content validity (i.e. the extent to which test content adequately samples the domain of interest), criterion validity (i.e. the extent to which a test correlates with another test of the same ability measured), or construct validity (i.e. the degree to which a test measures what it purports to measure).

A variety of approaches can be taken to determine construct validity, and these include the examination of "theoretical knowledge of the trait and ability being measured, knowledge of other related variables, hypothesis testing, and statements regarding the relationship of the test variable to a network of other variables that have been investigated" (Groth-Marnat, 2003, p. 22).

The present study set out to investigate the extent to which age and language aptitude influence long-term L2 learning outcomes, which is interpreted in terms of a

distinction between implicit and explicit knowledge. A series of tests were thus administered as hypothesised measures of these two types of linguistic knowledge. Several tests in the battery of tests of R. Ellis et al. (2009) were adopted in the present study, including the EI test, the TGJT, the UGJT, and the MKT. However, without making changes to the test materials, aspects including test administration and test modality (i.e. in aural or in written mode) were adjusted in the present study. Firstly, a further time pressure of repeating within 10 seconds for each stimulus sentence was imposed in the EI in addition to the original requirement that test takers heard the stimulus only once and in real time (Erlam, 2006; 2009). Secondly, test stimuli of the TGJT were presented aurally rather than in written form as in Loewen (2009) (see Section 3.4 for a description of these modifications). The present study also incorporated two other tests as proposed measures of implicit L2 knowledge, namely the WordM test and the cloze test. The WordM test was embedded in the TGJT and the cloze test was administered independently. In the end, a total number of six tests were used as language measures.

The validity of the tests described in R. Ellis et al. (2009) has been supported by individual validation studies (e.g. Erlam, 2006; Elder, 2009) and other studies that specifically address the construct validity of these tests (e.g. R. Ellis, 2005; R. Ellis et al. 2009; R. Ellis & Loewen, 2007; Zhang, 2015), which have shown that it was indeed possible to design tests that could access each type of knowledge independently. However, considering the modifications made to the EI and the TGJT, and the addition of the WordM and the cloze test as measures of implicit knowledge, the extent to which the data of the present study provides further evidence for the validity of the tests of implicit and explicit knowledge is examined.

With a view to validating these tests, multiple statistical analyses were conducted. First, the reliability of each test was checked by calculating the internal consistency estimates (see Section 3.5), and then the inter-correlations among tests were examined by doing a Pearson product-moment correlation analysis (see Section 4.3.1), followed by an exploratory factor analysis to detect the underlying constructs

of these tests (see Section 4.3.2) and a confirmatory factor analysis (see Section 4.3.3) to verify these constructs according to theory. Test scores of the NS participants were also consulted and compared to the scores of the NNS participants in order to learn about the construct validity of these tests. Based on the results of these analyses, and also by comparing and evaluating relevant previous findings, the construct validity of each test is discussed in detail in the following sections.

5.1.1 Tests of implicit knowledge

Oral elicited imitation test

The construct validity of the EI test as a measure of implicit knowledge is supported by the results of the present study. To begin with, the EI test had a statistically significant correlation with the TGJT, which was another hypothesised measure of the same type of knowledge ($r = .70$, $p < .001$). This means that these two tests were assessing knowledge that is related. In the meantime, the fact that the EI test was not significantly correlated with the MKT, a test that is generally agreed upon as a measure of explicit knowledge, shows that the EI test was assessing knowledge that was not related to that assessed by the MKT.

The lack of correlation between the EI test and the MKT contradicts the findings of Suzuki and DeKeyser (2015), in which metalinguistic knowledge was found to be a significant predictor of EI test. However, their EI test included a built-in word monitoring task (a word was presented for each sentence and the participants were required to press a key once they heard this word), which inevitably required the participants to focus on form to some extent. The absence of an opportunity to focus on form, however, as suggested by Erlam (2006), is perhaps the most important design feature for an EI test to measure implicit knowledge. Another design feature of Suzuki and DeKeyser's (2015) EI test which might have further enhanced the likelihood of focusing on form was to require the participants to correct ungrammatical sentences. Again, explicitly asking the participants to make

error corrections directed the attention of the participants to linguistic form and thus could lead to the involvement of explicit knowledge. Therefore, the correlation found between the EI test and the MKT by Suzuki and DeKeyser (2015) could be attributed to the fact that their EI test directed the attention of participants to form. Consequently, the fact that they found that their EI test accessed a similar type of knowledge with the MKT should be considered with circumspection.

The lack of correlation between the EI test and the MKT in the present study ($r = .18$, $p > .05$) not only provides evidence for the construct validity of these tests, but also underlines the vital role of test design and administration. As discussed earlier, due to differences in test design and administration, the EI test could be accessing different types of knowledge. To measure implicit knowledge, an EI test needs to have a primary focus on meaning (Erlam, 2006; 2009). In Erlam (2006; 2009), as in the present study, this focus on meaning was operationalised by having participants indicate whether they think belief statements are true or not. The premise is that in order to make such a decision, they have to comprehend the stimuli. The requirement that they make this belief choice also introduces some delay between the presentation of stimuli and repetition of it, which reduces the possibility of the involvement of rote memory. In test instructions, the participants should not be told that they will hear ungrammatical sentences, nor should they be instructed to make any corrections as in Suzuki and DeKeyser (2015). Instead, in the present study, the participants were told to "repeat in correct English" as suggested by Erlam (2006). Participants' correction of ungrammatical sentences was considered to be a likely indication of implicit knowledge. On top of these design features, an imposed time pressure whereby each stimulus was played only once and in real time would further enhance the likelihood of requiring participants to have to rely on implicit knowledge (Erlam, 2006; 2009). All of these aspects in test design and administration would potentially influence the validity of an EI test, and therefore should be considered thoroughly.

The results of the factor analyses lend further support to the construct validity

of the EI. The very high factor loading of .95 demonstrated a close relationship between this test and the hypothesised factor, implicit knowledge (see Section 4.3.3). This factor loading was even higher than those reported in R. Ellis and Loewen (2007; $\lambda = .87$) and Zhang (2015; $\lambda = .76$), which could be due to the more stringent time pressure. The stringent time pressure was the constraint of 10 seconds for sentence repetition that was imposed in addition to the time pressure of the original test (Erlam, 2006), that is, to hear each stimulus only once and in real time (see Section 3.4.1.2). Hypothetically, this reinforced time pressure would impose a heavier burden on the participants, and thus reduce the likelihood of resorting to explicit knowledge to monitor their sentence repetition. Based on the observations in the pilot study (see Section 3.6) and the main study, the fact that there were cases when some participants failed to complete their sentence repetition within the 10 seconds time frame showed that this was a further time pressure. Judging from the pattern of factor loadings in the confirmatory factor analysis (see Section 4.3.3), the results of the present study suggest that such an implementation of stringent time pressure together with meticulous test design and administration might have made the EI test a likely measure of implicit knowledge, and perhaps a better measure than the TGJT.

However, an alternative explanation could be that this higher loading is due to chance. Since different participants are involved in different studies, it is very likely that each study will obtain different factor loadings. On the other hand, judging from the magnitude of the factor loadings of the EI test as reported in the limited number of studies providing results of confirmatory factor analysis, it appears that this test repeatedly receives relatively high loadings (R. Ellis & Loewen, 2007; Zhang, 2015). This is an indication of the validity of this test as a primary measure of implicit knowledge (Erlam & Akakura, 2016; Granena, 2016). The present study provides another piece of empirical evidence in this regard.

Timed grammaticality judgement test (TGJT)

The construct validity of the TGJT as a measure of implicit knowledge is also supported by the results of the present study. As discussed earlier, some evidence is provided by the high correlation between the TGJT and the EI test ($r = .70$, $p < .001$), which shows that what was accessed by the TGJT was related to what was accessed by the EI test. Further evidence comes from the results of the confirmatory factor analysis, showing that the hypothesised common factor, implicit knowledge, was indeed accessed by the EI test and the TGJT ($\lambda = .95$ and .73, respectively). These findings corroborate previously reported confirmatory factor analysis results by R. Ellis and Loewen (2007), Bowles (2011) and Zhang (2015), all showing similar factor relations between implicit knowledge and the scores of the two tests.

The postulation that operationally manipulating time available to the participants of GJTs would tap into different types of knowledge is therefore supported (R. Ellis, 2005; Godfroid et al., 2015; Loewen, 2009). Hypothetically, the limited time allocated for making grammaticality judgements requires automatic processing and draws on implicit knowledge. Empirical evidence that attests to this assumption comes from making comparisons between the NS and NNS in their judgement accuracy of the grammatical and the ungrammatical sentences in the TGJT. The NS, whose knowledge is largely implicit, should be consistently accurate in their judgement irrespective of the grammaticality of the stimuli sentences. In contrast, the NNS, with arguably more inter-group variation in implicit knowledge, might be influenced by the grammaticality of the stimuli sentences. In other words, the construct validity of the TGJT as a measure of implicit knowledge is supported if the grammaticality of the stimuli sentences only influences the NNS but not the NS. The results of the present study show that this was indeed the case. The NS group scored similarly well in the judgement of the grammatical and the ungrammatical stimuli (see Table 4.3), and subsequent mixed models results confirmed that the grammaticality of stimuli sentences did not influence the judgement accuracy of the NS (see Section 4.4.1). Comparatively speaking, the NNS were much less accurate

in their judgement of the ungrammatical sentences (see Table 4.3), and the odds of making a correct judgement decreased by 91% when facing ungrammatical stimuli (see Section 4.4.1). This drastic change in judgement accuracy of the NNS could be caused by a lack of implicit knowledge, so that they were unable to make accurate judgements under time pressure. In turn, this means that the implementation of time pressure may have necessitated the NNS to rely on implicit knowledge, rather than explicit knowledge. Therefore, the construct validity of TGJT as a measure of implicit knowledge is supported.

Another source of evidence in support of the construct validity of the TGJT can be seen from the discrepancy in the test performance of the NNS in the TGJT and the UGJT. The NNS scored around 86% in the UGJT (see Table 4.4), but only 66% in the TGJT (see Table 4.3). Again, the inadequacy of implicit knowledge of the NNS could explain these results, and they therefore provide some evidence for the construct validity of the TGJT. Considering the fact that the majority of the NNS in the present study had been learning English in an instructed context for many years (see Table 4.63), it is expected that they probably had explicit knowledge of most of the 17 targeted grammar structures, including structures such as regular past tense, modal verbs, possessive –s, et al. Of course, this assumption does not apply to those who started their immersion in a naturalistic context (i.e. New Zealand) at a young age, for example, the young group who moved to New Zealand before 13 years old, because they did not receive explicit instruction in grammar. Assuming that the majority of the NNS had explicit knowledge of the target structures, but did not necessarily have the corresponding implicit knowledge, they would consequently perform better in the tests that allow for the use of explicit knowledge than those that primarily require implicit knowledge. The imbalance of test performance on the UGJT and the TGJT therefore, provides some evidence in this respect.

Another design feature that may have enhanced the likelihood of the TGJT tapping into implicit knowledge is its aural modality. Unlike R. Ellis and Loewen (2007), Loewen (2009) and Zhang (2015), the TGJT was aurally presented to the

participants in the present study. Overall, the factor loading of this aural TGJT (λ = .73) in comparison to those in previous studies where the same set of stimulus sentences was used in written format (λ = .68 in R. Ellis & Loewen, 2007; and λ = .47 in Zhang, 2015), suggested that the aural TGJT is perhaps a better measure of implicit knowledge than the written one. Several other studies that addressed the modality issue of the TGJT also had similar findings (Kim & Nam, 2017; Spada et al., 2015). For example, Spada et al. (2015) compared an aural and a written TGJT in an exploratory factor analysis. Their results showed a higher factor loading of the aural TGJT on the factor named implicit knowledge than the written one. Meanwhile, they also found a cross loading of the written TGJT on the factor named explicit knowledge, which led them to conclude that the written TGJT did not preclude the possibility of using explicit knowledge alongside implicit knowledge. Similarly, Kim and Nam (2017) found a higher factor loading of the aural TGJT than the written one on a factor named weaker implicit knowledge. From these results, it seems that aural TGJTs are mainly related to implicit knowledge, while written TGJTs might be relevant to both implicit and explicit knowledge.

However, the conclusions of Spada et al. (2015) and Kim and Nam (2017) are somewhat weakened due to limitations in data analysis methods. Had the data of Spada et al. (2015) been submitted to confirmatory factor analysis, the extent to which their aural and written TGJT measure the same or different constructs could have been examined in a more robust manner. Unfortunately, only the results from an exploratory factor analysis were reported in their study. Likewise, there are some methodological limitations in Kim and Nam's (2017) study. The major problem is model misspecification, especially relating to the number of indicators (i.e. tests) that each hypothesised factor should have. From a statistical perspective, a confirmatory factor analysis of models with three latent factors would require at least three manifest indicators (i.e. three different test scores) for each factor (Brown, 205); but this requirement was not met in the study of Kim and Nam (2017). Specifically, for the factor named explicit knowledge, only one test (i.e. the MKT) was specified as

its indicator. This kind of model specification is cautioned against by statisticians. As a result, this flaw in model specification undermines the validity of the conclusions of Kim and Nam (2017).

Due to the fact that a written TGJT was not included in the present study, a direct comparison of the aural TGJT and a written TGJT could not be made. However, the analysis of results provides some evidence for the aural TGJT as a measure of implicit knowledge. However, one might raise the concern that the alteration to the aural mode of the TGJT would bring method effects, meaning that tests that adopt the same measurement modality (i.e. aural mode of the EI test and the TGJT) are correlated and load on the same factor (Spada et al., 2015). With the benefit of confirmatory factor analysis, the possibility of this method effect was addressed in the present study. If the modality of these two tests had indeed had an impact, there would have been an error covariance between the EI test and the TGJT, showing that what was not explained in the data by the common factor was explained by the shared testing modality. However, as indicated by the results (see Figure 4.3), such an error covariance was not present. That is to say, just like the previously used written TGJT, the aural TGJT also served as a good measure of implicit knowledge.

A possible explanation for the involvement of implicit knowledge in the aural TGJT can be attributed to the heavier processing burden imposed on participants. Research in sentence processing has shown that it is more challenging to process aural language than written text (McDonald, 2000), probably because verbal information presented in the auditory mode is processed by a short-term system that has limited capacity (Penney, 1989). This means that it is particularly difficult to attend to both form and meaning when the input is presented aurally (Wong, 2001), and the less attention paid to form, the less likely it is that explicit knowledge is involved. In other words, the larger processing demand would limit the accessibility of explicit knowledge when test materials are presented aurally (Johnson, 1992). Based on these studies, it can be concluded that in comparison to written TGJTs,

aural TGJTs are more likely to force participants to draw on implicit knowledge. However, the potential of the aural TGJT as a measure of implicit knowledge still warrants further research.

Word monitoring test

The construct validity of the WordM test as a measure of implicit knowledge is not supported by the results of the present research. In light of previous research that used this test as a measure of implicit knowledge (Granena, 2012; 2013a; Suzuki & DeKeyser, 2015), the present study built a word monitoring test into the aural TGJT. However, it was found that the WordM test shared very little variance with the TGJT and the EI test (as low as 6%), which means that the knowledge being used in the WordM test was different from that used in the EI test and the TGJT (see Section 4.3.1). The results of the factor analyses (see Sections 4.3.2 and 4.3.3) further suggested that this test was not connected to either implicit or explicit knowledge, or more conservatively, not connected to the constructs measured by other tests used in the present study. This is an unexpected finding since some researchers have argued that the WordM test is a measure of implicit knowledge superior to the EI (Suzuki & DeKeyser, 2015; Suzuki, 2017; Vafaee et al., 2017). These researchers maintain that the WordM test is an indirect and online measure of grammatical sensitivity which avoids requiring participants to make grammaticality judgements. A delay in reaction time when ungrammatical sentences are encountered is seen as evidence to suggest sensitivity to grammatical violations, which indirectly implies that implicit knowledge is activated.

The discrepancy between the results of the present study and previous research could be attributed to differences in the overall designs across studies, especially in terms of other tests involved alongside the WordM test. Studies that favour the use of WordM test as a measure of implicit knowledge have also used other psycholinguistic tests such as the serial reaction time test (Granena, 2012; 2013b) and the self-paced reading test (Suzuki, 2017). These tests produce reaction time

data (in milliseconds), which is different from the test scores of a TGJT or an EI test. It is possible that the assembling of these psycholinguistic tests in factor analyses is because of their shared format of data (i.e. reaction time data), which is usually skewed and very different from the scores obtained from other tests. In addition, the relatively low factor loadings of these psycholinguistic measures as reported by previous researchers might not be robust enough to demonstrate their relatedness to the construct of implicit knowledge (see Section 5.1.3 for more discussion). Clearly, further research is warranted to investigate the extent to which the WordM test measures implicit knowledge.

5.1.2 Tests of explicit knowledge

Untimed grammaticality judgement test

The construct validity of the UGJT as a measure of explicit knowledge is supported by the findings of the present research. Some evidence in this respect comes from the observation that there was a significant correlation between this test and another test that was designed as a measure of the same type of knowledge, that is, the MKT (see Section 4.3.1). The correlation of .63 ($p < .001$) shows that what was accessed by the UGJT was related to that which was accessed by the MKT. However, the correlation analysis results seemed to have indicated somewhat contradictory findings because significant correlations were also found between the UGJT and the two tests of implicit knowledge, namely the EI test ($r = .64$, $p < .001$) and the TGJT ($r = .49$, $p < .001$). This means that what was measured by the UGJT was probably related to implicit knowledge too. In fact, it is possible to complete this test using implicit knowledge although the test condition is designed to favour the use of explicit knowledge. For example, it can be seen that the NS, whose knowledge is largely implicit, scored around 94% in this test. Presumably some NNS with a high level of proficiency were also able to complete the UGJT using their implicit knowledge, which could explain the correlations found between the UGJT

and the tests of implicit knowledge.

Nonetheless, further evidence that lends support to the construct validity of the UGJT comes from the results of the factor analyses (see Sections 4.3.2 and 4.3.3). First, some direct evidence was provided to show that what was accessed by the UGJT was different from what was measured by the tests of implicit knowledge because these tests loaded on two different factors. Moreover, there was some evidence that explicit knowledge was indeed involved in the UGJT, judging from the significant and high factor loading of .88 (see Section 4.3.3 and Figure 4.3). Therefore, it is concluded that the UGJT is a likely measure of explicit knowledge, rather than implicit knowledge.

The overall good performance of the NNS in the UGJT (mean score 85.7%) in comparison to only 60.5% in the TGJT is some evidence that explicit knowledge can be accessed without time pressure. Further evidence that lends support to this comes from the finding that the grammaticality of the stimuli sentences does not affect the judgement accuracy of the NNS in the UGJT (see Section 4.4.2). Because the participants had sufficient time to reflect on their grammatical knowledge, the ungrammatical sentences that they were unable to identify in the TGJT were accurately judged in the UGJT.

The investigation of the relationship between the degree of certainty and the judgement accuracy of the NNS shows that there is a positive linear relationship between the two (see Section 4.5 and Figure 4.7). This linear relationship was found even after controlling for individual variations in the level of certainty across test items (i.e. random effect of subject and item) in the mixed logit model CertaintyNNS1. This finding contradicts the hypothesis that explicit knowledge is often imprecise and inaccurate (Sorace, 1985). Some participants may be very likely to express confidence in their explicit knowledge, especially if the target language grammar has been explicitly taught and learnt. In the UGJT, a participant may have identified the error, and reflected on the corresponding grammatical rule, and thus placed considerable confidence in his or her judgement. If explicit knowledge was

indeed used in the UGJT, this finding indicates that the participants were generally confident with their use of explicit rules. However, other participants may be more prone to express confidence when they draw on their implicit knowledge if this type of knowledge is used in this test. If this is the case, a high level of certainty manifests the inherent nature of implicit knowledge, that it is highly systematic and accurate (R. Ellis, 2005; 2009). Thus, since it is difficult to rule out the possibility of using implicit knowledge in the UGJT, it is difficult to decide on what basis (implicit or explicit knowledge) the participant expresses their level of certainty. As R. Ellis (2005) noted, the use of certainty to infer which type of knowledge is used in the UGJT needs to be treated with circumspection.

The self-reported use of rule rather than intuition by the participants in the UGJT shows that it was very likely that this test imposed a primary reliance on explicit knowledge for the NNS, which lends support to the construct validity of this test to some extent. Of the 86 NNS participants, 72% reported that they relied on grammar rules and 15% said that they mainly relied on intuition. The remaining 13% of the NNS participants reported that they mainly relied on intuition but also used their knowledge of grammar to analyse complicated sentences with more than one clause. This data seems to imply that the NNS considered that they primarily used explicit knowledge in this test. However, it should be noted that self-reported data like this could be lacking in reliability. If implicit knowledge was indeed being used by some of the NNS participants as they claimed, they probably would have performed better than those who mainly relied on explicit knowledge because implicit knowledge is more likely to be accurate. However, the one way ANOVA analysis results indicated no significant differences between those who relied on feel, on rule or on both feel and rule (see Section 4.5; $F(2, 81) = .30$, $p = .74$). In fact, those who relied on rule ($n = 61$) obtained the highest mean score in this test ($M = 58.90$, $SD = 6.10$). These results contradict the assumption that those who relied on feel would score highest in this test. In turn, this finding suggests that self-reported data is not very reliable. However, it is also possible that the small within group

variation of the NNS caused difficulty in examining the relationship between judgement accuracy and the reported use of rule or feel since all NNS performed rather well in this test.

Metalinguistic knowledge test

Metalinguistic knowledge refers to knowledge that is explicit and analytical in nature. It is comprised of knowledge of declarative facts, including rules or fragments of information, regarding a language (Elder, 2009). To test metalinguistic knowledge, a test that involved metalanguage, that is, the technical or semitechnical terminologies used to describe grammar (James & Garrett, 1992), was used in the present study (R. Ellis, 2005; Elder, 2009). The construct validity of the MKT as a measure of explicit knowledge is supported by the present study since this test was found to have had a significant correlation with another measure of explicit knowledge (i.e. the UGJT, $r = .56$, $p < .001$) while it had very low correlations with the hypothesised measures of implicit knowledge (i.e. the EI test, the TGJT and the WordM test; $r = .18, .07, .08$, respectively). This means that what was accessed by the MKT was different from that which was accessed by the hypothesised tests of implicit knowledge.

Further evidence that lends support to the construct validity of the MKT comes from the confirmatory factor analysis results (see Section 4.3.3), in which this test loaded on the hypothesised factor explicit knowledge. However, this test received the lowest factor loading ($\lambda = .33$). The reason for this could be that one's command of metalinguistic knowledge is independent of explicit knowledge (Clapham, 2001; R. Ellis, 2004). In particular, individuals vary to a large extent in their command of metalanguage. This means that one may be able to display the explicit knowledge of grammatical rules without knowing the metalanguage to technically verbalise these rules. Therefore, it is not surprising to see a rather weak connection (i.e. factor loading of .33) between the MKT that involved much metalanguage and the hypothesised factor of explicit knowledge.

The asymmetry between one's metalinguistic knowledge and explicit knowledge is evident given the results of the present study. On the one hand, the NNS scored 85.7% on average in the UGJT, which indicated a relatively high level of explicit knowledge. On the other hand, they scored only 61.7% in the MKT, indicating a lower level of metalinguistic knowledge. Although the participants were not asked to verbalise rules in grammatical terms in the MKT, the understanding of metalanguage was an essential part of test completion. As a consequence of a lack of metalinguistic knowledge, their performance in this test was not as good as in the UGJT, and therefore the hypothesised factor explicit knowledge could only account for a limited amount of variance in the test scores.

Some further evidence that lends support to the construct validity of the MKT comes from the observation that the NNS group outperformed the NS group in this test. The statistically significant between-group difference ($t(104) = -2.32$, $p = .02$; see Section 4.2) showed that the NNS participants had a higher level of metalinguistic knowledge than the NS. This difference could be attributed to the types of learning that these two groups of participants engaged in. The NS participants primarily acquired their knowledge implicitly and thus would have limited explicit knowledge of English grammar. In comparison, the majority of the NNS participants ($n = 70$) learnt some English grammar knowledge in EFL classrooms, where English was taught in an explicit way. Specifically, as reported by these NNS, English grammar was taught using a deductive approach, which involved much explanation of grammar rules using metalanguage (see Table 4.64). As a consequence, it was expected that the NNS at a group level would have more metalinguistic knowledge than the NS. The finding that the NNS indeed performed better than the NS in the MKT therefore lends support to the validity of this test as a measure of explicit metalinguistic knowledge.

It can be seen from the self-reported data that deductive instruction of grammar was the prevalent practice in China. This is understandable because English teaching in China tends to be teacher-dominated and exam-oriented, rather than

student-centred and communicative-oriented (G. Hu, 2005; Yu & Wang, 2009). Under such an instructed learning context, the NNS were explicitly taught the grammar rules first and then required to remember these rules. Both rule explanation and memorisation would inevitably involve the use and understanding of metalanguage, which in turn potentially led to the development of metalinguistic knowledge. However, it seems that despite this kind of deductive instruction that the NNS were exposed to, there was a relatively low level of metalinguistic knowledge developed in the end. This finding contradicts previous research findings that the learning environment and specifically the instruction method (i.e. communicative approach vs. grammar-focused approach) determine learners' access to metalinguistic knowledge (Elder & Manwaring, 2004; Renou, 2001). In the present study, the overall metalinguistic knowledge of the NNS at group level (61.7%) was less than expected. It is possible that after a rather long period of residence in the L2 country, the NNS participants found it hard to access their metalinguistic knowledge. At the time of test completion, some participants told the researcher that they had memories of learning grammar rules and all the relevant metalanguage, but it was very difficult for them to recall and to identify accurate explanations. It is also possible that as learners develop more implicit knowledge, their dependence on metalinguistic knowledge gradually decreases (P. S. Green & Hecht, 1992). As a consequence, when asked to utilise their metalinguistic knowledge, the NNS participants received relatively low scores.

Although there is some evidence attesting to the construct validity of the MKT, the findings of the present study also suggest that the test materials of the MKT need to be prepared carefully, especially if another test of explicit knowledge is simultaneously administered. This suggestion is made on the basis of the observed significant error covariance between the MKT and the UGJT in the confirmatory factor analysis ($r = .60$, $p < .001$). The presence of this error covariance indicated that the variations and covariations of these two tests that were not accounted for by the factor explicit knowledge were caused by another exogenous factor.

Hypothetically, the common measurement method, that is, the written modality of these two tests, could have influenced the performance of the NNS to some extent. In addition, the content of the test items in these two tests could be another contributing factor. Both of these tests targeted the same 17 grammar structures and therefore the stimulus sentences were similar. The participants did the MKT right after they finished the UGJT, which means that they read sentences of similar content consecutively. This overlap in test content and modality could probably explain the relation of correlated error variance between these two tests. Therefore, to better measure metalinguistic knowledge, and also explicit knowledge, future research could benefit from diversifying test materials.

Cloze test

The cloze test was intended as a measure of implicit knowledge, but it was found that this test was primarily testing explicit L2 knowledge in the present study. Some evidence in this respect is provided by the correlation analysis (see Section 4.3.1). The fact that the cloze test had stronger correlations with the tests of explicit knowledge, that is, $r = .63$ with the UGJT ($p < .001$) and $r = .36$ with the MKT ($p < .001$), than with the tests of implicit knowledge, that is, $r = .53$ with the EI ($p < .001$) and $r = .33$ with the TGJT ($p < .001$), shows that what was accessed by the cloze test was more related to that accessed by the tests of explicit knowledge. On the other hand, these results also suggested that implicit knowledge might also have been used in the completion of this test.

However, the factor analyses results (see Sections 4.3.2 and 4.3.3) further support the association between the cloze test and explicit knowledge ($\lambda = .68$, $p < .001$). Both the exploratory and the confirmatory factor analysis showed that it was explicit knowledge rather than implicit knowledge that was primarily involved in the completion of this test. Again, as it is impossible to devise pure measures of implicit or explicit knowledge (R. Ellis, 2005), it is concluded that the factor analysis results demonstrated a stronger connection between explicit knowledge and the cloze test.

The performance of the NS participants in the cloze test provides further evidence that this test potentially accessed knowledge other than implicit knowledge. Arguably the NS participants would use their implicit knowledge to complete this test and would have the ability to obtain high scores. However, their average performance of 66.6% does not demonstrate a high level of accuracy in this test. The mediocre performance of the NS group could be an indication that this test was not exclusively measuring implicit knowledge. If the cloze test was a good measure of implicit knowledge, the NS should have demonstrated similarly high average scores as they did in the TGJT and the EI. In other words, despite the fact that the NS participants had implicit knowledge of collocations, some other knowledge, purportedly explicit analytical knowledge, was also required by this test. Previous research has also shown that cloze tests measure a range of knowledge, including grammatical, vocabulary, orthographic, semantic, and pragmatic knowledge (Hulstijn, 2010). This means that explicit knowledge, especially the extent to which ascertaining which words can complete missing gaps in sentences, requires a knowledge of grammar (e.g. third-person singular).

The above discussed results indicate that rather than collocational knowledge, the cloze test was in fact tapping into more grammatical knowledge. This is probably a result of test content since this test did not directly test collocational structures such as “do homework” and “make the bed”. Rather, the cloze test examined the use of constructions such as “not… but …” and “while others” (see Questions 3 and 6, Appendix 3). These constructions differ from collocational structures such as “do homework” and “make the bed” in the sense that they could be learnt explicitly, while the latter are more likely to be acquired implicitly in an L2 environment. In other words, although the cloze test was intended as a measure of implicit knowledge, it turned out that implicit knowledge was not necessarily required to complete this test. What was actually required could be knowledge that the NNS learnt explicitly in an English classroom in China, where the use of constructions such as “not… but …” and “while others” was taught. The self-reported learning

experience of the NNS provide some evidence in this respect because explicit deductive teaching seemed to be the main teaching method of teachers in China (see Table 4.64), which might have enabled the development of the explicit knowledge that was required by the cloze test. As a result, in statistical analyses, scores of the cloze test clustered together with tests that accessed a similar type of knowledge (i.e. the UGJT and the MKT).

In addition, the fact that the cloze test was in a format that was very similar to the multiple-choice cloze exercise that is typically given in English classes in China might have enhanced the likelihood of this test tapping explicit grammatical knowledge. The researcher herself and the NNS participants who had experienced learning English in China were familiar with this test format and had done many such exercises in preparation for the national college entrance examination. The common approach to dealing with this kind of cloze exercise was to adopt an analytical method by using the grammatical features of the sentences to infer the words, and then select from the multiple choices. So when given a similar cloze test in the present study, some NNS participants may have habitually adopted an analytical approach. For example, in sentence 14 of the cloze test (Please take the bandages off my head, _____feels sorer and _____.), "feels" in singular form gave the participants a clue that they need to put a subject in third person singular in the gap. Consequently, the use of such an analytical approach might have led to the involvement of a high degree of awareness of grammatical rules and thus a primary dependence on explicit knowledge.

The use of such an analytical approach also potentially brought about an attention to form, which again might have triggered the use of explicit knowledge. Moreover, since the cloze test was untimed, the participants could take as long as they wished to analyse each sentence and consequently the chances of explicit knowledge dominating the test completion process were increased. A high degree of awareness, attention to form, and the absence of time pressure are all operational criteria that encourage the use of explicit knowledge (R. Ellis, 2005; 2009). The

cloze test happened to create a condition that embraces all these criteria, and this may have necessitated a primary dependence on explicit knowledge of the NNS, rather than implicit knowledge. As a consequence, both the correlation analysis and the confirmatory factor analysis indicated that this test was connected to explicit knowledge.

An alternative explanation for the primary use of explicit knowledge by the NNS could be a lack of implicit knowledge. Evidence in support of this assumption comes from the observation that the NNS spent an average of 20 minutes in this test while the NS spent an average of only 10 minutes. The NS spent less time because implicit knowledge was at their disposal although they might not have relied exclusively on this type of knowledge, while the NNS were unable to fill in the gaps by relying on the same type of knowledge. To compensate for the deficiency of implicit knowledge, the NNS had to resort to explicit knowledge to analyse the sentences and therefore they spent twice as long as the NS in this test.

5.1.3 The implicit-explicit model

The analysed results of the present study, especially those from factor analyses, demonstrate that L2 knowledge can be empirically separated as implicit and explicit knowledge. In other words, language tests such as those devised by R. Ellis (2005) and R. Ellis et al. (2009), can be devised to tap into implicit or explicit knowledge separately. With some modifications made to the tests of R. Ellis et al. (2009), together with rigorous data analysis methods, the construct validity of these hypothesised tests of implicit and explicit knowledge is ascertained. In the present study, implicit and explicit knowledge were hypothesised as independent constructs, and multiple tests were employed to gauge the multiple dimensions that these constructs entail. Overall, the current research findings lend support to the construct validity of the six language tests as tests of implicit and explicit L2 knowledge.

The successful dissociation of implicit and explicit knowledge owes much to

the data analysis method employed, namely exploratory and confirmatory factor analysis. The use of this data analysis method has several advantages. To start with, the number of latent variables (i.e. the factors) that account for the variance and covariance of the scores of a set of tests can be determined using a data-driven approach (i.e. exploratory factor analysis) and/or a theory-driven approach (i.e. confirmatory factor analysis). The relation between the latent variables can also be examined and evaluated during this process. On top of this, another advantage of factor analysis is that it is based on the common factor model, which assumes error variance that is not explained by the common factor. Such an assumption of error variance brings benefits to applied research because accurate measurement of the intended construct is rarely the case. Therefore, statistical methods that recognise the existence of error variance are more likely to make estimations closer to population values, and therefore provide better statistical explanation and estimation (Brown, 2015).

The fact that the same implicit-explicit model was achieved by the data-driven exploratory factor analysis and the theory-driven confirmatory factor analysis attests to the distinction between implicit and explicit knowledge with the use of the six language measures (see Sections 4.3.2 and 4.3.3). As explained earlier, by adopting a data-driven approach, the intent of the exploratory factor analysis was to discover the appropriate number of common factors that reasonably explained the variation and covariation in the data of the language tests. The analysis of results indicated a two-factor model, although no a priori specifications were made in regard to the number of factors to be extracted. The size and pattern of factor loadings of this two-factor model were also reasonable, showing that the hypothesised measures of the same construct loaded on the same factor.

The decision of which factor extraction and rotation method to use in factor analysis needs to be considered in order to obtain convincing results. However, it seems that previous research has neglected these aspects to some extent and has not justified choices of factor extraction and rotation method. For example, principal

component analysis has been used as the factor analysis in some studies (R. Ellis, 2005; Granena, 2012; Gutiérrez, 2013; Zhang, 2015). However, the major weakness of principal component analysis is that it is not based on the common factor model and thus it does not account for error variance. Statisticians have pointed out that it is more appropriate to use principal component analysis for the reduction of a large set of data to a more manageable size, rather than to infer the number of latent factors that elucidate variations in the data (Brown, 2015; Fabrigar, Wegener, MacCallum, & Strahan, 1999). In lieu of principal component analysis, a principal axis factoring method that was based on the common factor model was used for model estimation in the present study. Similarly, the choice of factor rotation method may also influence the interpretation of the results. Some previous research has used orthogonal rotation methods that do not allow for correlation between factors, such as Varimax (e.g. Granena, 2012), but without giving reasons for this particular choice. Considering the possible relationship between factors in the present study (i.e. implicit and explicit knowledge), an oblique rotation method that assumes inter-factor correlation was used. Such a factor rotation method was in line with the theoretical hypothesis.

The results of the exploratory factor analysis showed that the data of the six language tests could be accounted for by two factors, hypothetically implicit and explicit knowledge. In other words, such results demonstrated that these tests measured two different constructs. In line with the correlation analysis results, the EI test and the TGJT loaded on one factor, whereas the UGJT, the MKT, and the cloze test loaded on the other (see Section 4.3.2). The cross loading of the UGJT on the factor named implicit knowledge reflected the unintended activation of this kind of knowledge in the completion of this test. This finding is consistent with results from previous studies (R. Ellis, 2005; Han & Ellis, 1998). The heavier factor loading of the cloze test corroborates the results of the correlation analysis, indicating a closer relation between this test and explicit knowledge. However, a finding that contradicts previous research results (Granena, 2012; Suzuki & DeKeyser, 2015;

Vafaee et al., 2017) emerged from the data analysis. The WordM test, a test that was intended as a measure of implicit knowledge, was not identified as an indicator of either implicit or explicit knowledge.

The results of the exploratory factor analysis were subsequently considered empirical evidence for the model specifications of the confirmatory factor analysis. That is, the structural pattern that emerged from the data was used in conjunction with theory to specify an a priori measurement model and a correlated factor relationship (i.e. model CFA1, Figure 4.2). Unfortunately, this initial run of factor analysis failed to meet the criteria for a good-fit model. Therefore, a modified model CFA2 (see Figure 4.3) was constructed and it was found that this model had a satisfactory level of model fit. During this process of model modification, the data of the WordM test was discarded, and an error covariance between the UGJT and the MKT was added. The analysis of results justified such modifications made to the initial model and suggested that the purpose of using this set of tests as separate measures of implicit and explicit knowledge was justified.

The results from the confirmatory factor analysis in the present study support an implicit-explicit factor model that corroborates those reported by R. Ellis and Loewen (2007), Bowles (2011) and Zhang (2015). The relationship between the two factors is described by the inter-factor correlation coefficient, which provides information about the discriminant validity, that is, the extent to which the hypothesised constructs are distinct. The estimated inter-factor correlation coefficient in the present study (r = .78) is lower than Bowles (2011, r = .87) and Zhang (2015, r = .86) but higher than R. Ellis & Loewen (2007, r = .56). In fact, all the above mentioned previous studies have been criticized from a statistical perspective by Vafaee et al. (2017), because no attempts were made to test a rival model (i.e. a one factor model). Vafaee et al. (2017) further argued that if a one factor model fits the data as well as a two-factor model, the construct validity of the two-factor model could not be supported. This is a fair criticism since factor correlations higher than .80 were reported by Bowles (2011) and Zhang (2015), and

these high correlations indicate poor discriminant validity (Brown, 2015). In situations where high factor correlations are found, the common strategy is to respecify a single factor model and then compare the model fit of the two models (Brown, 2015). Following this practice, a restrained model that fixed the inter-factor correlation to one was tested and compared to the model CFA2 in the present study (see Section 4.3.3). This restrained model was equivalent to a model that has only one factor accounting for all of the variations and covariations of the six language tests. The Chi square difference test results suggested that the restrained model significantly reduced the overall model fit. This meant that the hypothesised two factors measured different constructs and a distinction should be made. This result is quite expected since previous data-driven exploratory factor analysis results suggested a two-factor solution as well. Nonetheless, by empirically testing the discriminant validity of the hypothesised two factors, direct evidence is provided to support the two-factor model of implicit and explicit knowledge.

Another aspect of the results of the confirmatory factor analysis that should be attended to is the interpretability of the parameter estimates (i.e. the factor loadings). However, previous research put primary emphasis on reporting the overall model fit and did not address this aspect sufficiently. Adequate evaluations of these parameter estimates are essential, since goodness of fit indices alone can be misleading (Brown, 2015). From a statistical perspective, the direction, significance and magnitude of the indicators should be checked. A common problem in confirmatory factor analysis is the presence of *Haywood cases*, such as standardised factor loadings that exceed 1.0 and negative error variances, both of which are indicators of model misspecification. The significance of the factor loadings in the model is also important, as non-significant factor loadings indicate that the observed measure is not related to its purported latent factor (Brown, 2015). Similarly, factor loadings should demonstrate substantive magnitude, which signifies whether the observed measures are salient indicators of the latent constructs. The commonly used .40 and above to determine the importance of indicators has been viewed as “too liberal” in studies that use sum

test scores (Brown, 2015, p. 115).

Following an evaluation of the parameter estimates of the present study, it is concluded that the model CFA2 is justified (see Figure 4.3), with all the language tests showing sizable, significant, and meaningful factor loadings. In addition, the higher-order factor model CFA3 (Figure 4.4) also shows that the variance of the two factors, implicit and explicit knowledge, can be accounted for by one common factor named L2 knowledge. This higher factor model lends further support to the validity of the language measures used in the present study. The only test that was not substantively accounted for by the latent factor is the cloze test, where a significant error covariance with the UGJT was present. However, it is unclear whether previous researchers conducted a similar process of model evaluation. Judging from the information provided in their paper, the implicit-explicit model of R. Ellis and Loewen (2007) seems reasonable. In comparison, the model presented by Bowles (2011) has an out-of-range factor loading of 1.02 and a relatively low factor loading of .44; and similarly low factor loadings of .42 and .47 are found in Zhang's (2015) model. Other than this information, a more thorough description of the model parameter estimates and model evaluation have not been reported. Considering the results of the present study and previous research, there is some support for the implicit-explicit model, but research adopting the same analytical method is still warranted to examine the validity of this two-factor model.

5.1.4 Conceptual issues regarding the implicit-explicit model

Evidence presented in the previous section demonstrates that the current research findings support the dichotomous view of L2 knowledge comprising implicit and explicit knowledge. However, this viewpoint of L2 knowledge has received critique at a theoretical level. A number of researchers have argued for automatized explicit knowledge, a type of explicit knowledge that has been automatized to the extent that it can be drawn on under time pressure (DeKeyser,

2003; Rebuschat, 2013; Suzuki & DeKeyser, 2015; Suzuki, 2017; Vafaee et al., 2017). In their opinion, the manipulation of time pressure resulting in speeded tests, such as the TGJT, does not result in a qualitative change in the type of knowledge being used in comparison to that used in an untimed test. To support their argument, Kachinske and Vafaee (2014), as cited in Vafaee et al. (2017), reanalysed the data of R. Ellis and Loewen (2007) and showed that a one-factor model fits the data as well as the two-factor model. Moreover, Vafaee et al. (2017) found that the TGJT, the UGJT and the MKT loaded on the same factor in a confirmatory factor analysis. Suzuki (2017) also provided similar evidence, showing that the loadings of the time pressured tests (an aural TGJT, a visual TGJT, and a simple performance-oriented test) loaded on a separate factor from other psycholinguistic reaction time measures (a visual word test, a word monitoring test, and a self-paced reading test). Based on this evidence, they have argued that time pressured tests, such as the EI test and the TGJT, are merely accessing automatized explicit knowledge, rather than implicit knowledge.

However, a note of caution is due here regarding whether the above listed findings could in fact be seen as evidence in support of a construct which we could label as automatized explicit knowledge since there are a number of uncertainties. To start with, there is a lack of information about model evaluation. It is unclear how Kachinske and Vafaee (2014) arrived at their conclusion since their paper was a conference paper that the public do not have access to. As stated earlier, relying solely on general fit indices of confirmatory factor analysis models may lead to inappropriate conclusions. A critical step to ensure the validity of any confirmatory factor analysis models is rigorous model evaluation. Unfortunately, this pivotal step appears to have been neglected by Vafaee et al. (2017). In their study, they included two psycholinguistic tests (a self-paced reading test and a word monitoring test) alongside a timed GJT, an untimed GJT and a MKT. They then compared a large number of confirmatory factor analysis models and arrived at the best fit model of implicit-explicit knowledge, where the two psycholinguistic tests were indicators of

implicit knowledge, and the rest indicators of explicit knowledge. Despite the good overall model fit they have reported in reference to various fit indices, one obvious issue in their model is the presence of a Heywood case, that is, an out-of-range factor loading of 1.34 (normally it should be less than 1) of the indicator UGJT and its negative error variance of -.80. The existence of such a case is very likely a result of model misspecification (Brown, 2015). Had this model been thoroughly evaluated by Vafaee et al. (2017), a different conclusion regarding its well-formedness would have been made. This Heywood case also reduces the power of their model, which in turn raises some concerns regarding their conclusions made on the basis of this model.

Another source of uncertainty in the proposed model of automatized explicit knowledge and implicit knowledge is the rather low factor loadings of the indicators (Vafaee et al., 2017). Perhaps due to the Heywood case, the MKT and TGJT test received factor loadings of .33 and .36 respectively, both of which are too low to conclude that these measures are good representations of the latent factor named explicit knowledge. Similarly, the factor loadings of their two psycholinguistic knowledge tests as measures of implicit knowledge were not high in magnitude (λ = .42 and .58). In confirmatory factor analysis models without cross-loading indicators, such factor loadings represent the correlation coefficients between the indicators and the factors. Thus, an estimation of the variance explained in the indicators can be obtained by squaring their loadings. A loading that is equivalent to .70 can roughly explain 50% of the variance of an indicator, and a loading of .40 can only explain 16% of the variance. The proportion in the indicators that are not explained therefore becomes very large (e.g. 1-16% = 84%). That is to say, in situations where the aim is to validate certain tests as reliable measures of hypothesised constructs, one would want to obtain factor loadings that are high in magnitude according to which the reliability of these tests could be justified.

Such relatively low factor loadings of the psycholinguistic tests as measures of implicit knowledge were similarly reported in Suzuki (2017). In his study, he

included an aural TGJT, a written TGJT, and a simple performance-oriented test as measures of automatized explicit knowledge. Three psycholinguistic measures, including a visual word test, a word monitoring test, and a self-paced reading test were employed as measures of implicit knowledge. Suzuki (2017) then tested several confirmatory factor analysis models and concluded that the three psycholinguistic measures measured implicit knowledge, while the rest measured automatized explicit knowledge. However, in his model using the data of 99 second language learners, the factor loadings of the psycholinguistic tests were all non-significant and relatively low (λ = .48 for the visual word test, .10 for the word monitoring test and .28 for the self-paced reading test); while the factor loadings of the tests of automatized explicit knowledge were all significant and much larger in magnitude (λ = .65; .80; .85). These nonsignificant low factor loadings probably indicate the irrelevance of the indicators to the hypothesised construct of implicit knowledge. Suzuki (2017) then split the entire group into two subgroups according to length of residence in the L2 country and conducted confirmatory factor analyses again. This resulted in reductions in sample sizes (n = 47 and n = 52), both of which are below the required number of samples for models with two hypothesised factors and six indicators (Brown, 2015; MacCallum et al., 1996). The results of using the data from different groups indicated somewhat different patterns: a one-factor model fitted the data of the short LoR group and a two-factor model fitted the data of the long LoR group. Suzuki's (2017) interpretation of these results was that the two types of knowledge were more distinct for the long-LoR group while the short-LoR group had a heavier reliance on automatized explicit knowledge. However, since both the short and long LoR group were L2 learners (i.e. both are from the same population), a reasonably good two-factor model should have the ability to demonstrate measurement invariance, that is, the extent to which the same model pertains to different samples of the same population. This measurement invariance, however, is not supported by the distinct models reported by Suzuki (2017). The inability to demonstrate such measurement invariance might be a reflection of the lack of

validity of the two-factor model as proposed by Suzuki (2017).

Suzuki (2017) went on to demonstrate the validity of the hypothesised constructs by conducting several multi-trait multi-model (MTMM) confirmatory factor analyses. However, the extent to which his research design falls in the scope of multi-trait multi-model analysis is not evident. The use of such analysis generally requires each trait (i.e. latent factor) to be measured by each method, and a minimal number of three traits and three methods are generally used (Campbell & Fiske, 1959; Kenny & Kashy, 1992). In Suzuki's (2017) study, each of the two traits was only measured by one method rather than by each of the two methods. Therefore, although the MTMM models showed an overall good model fit, some caution should be taken in the interpretation of the results generated.

Moreover, in models where online psycholinguistic tests were specified as measures of implicit knowledge, the between factor correlations were relatively low (Suzuki, 2017; Vafaee et al., 2017). Suzuki (2017) obtained a factor correlation of .32 from his MTMM model, and Vafaee et al. (2017) reported a correlation of .26. Such low correlations undoubtedly attest to the discriminant validity of the hypothesised constructs but are difficult to understand especially since research findings from different disciplines generally agree that implicit knowledge (or automatized explicit knowledge) and explicit knowledge are of mutual influence (N. C. Ellis, 2005). The non-significant correlations reported by the above mentioned researchers therefore, seem to imply that implicit and explicit knowledge are so distinct that there is minimal correlation between them. Automatized explicit knowledge, if placed on a continuum of explicit to implicit, should be closer to the implicit end, and therefore it should be more likely to correlate with implicit knowledge. However, the non-significant and relatively low correlation reported in Suzuki (2017) seems to provide counter evidence in this respect. Nonetheless, as Suzuki acknowledged himself, it is empirically difficult to develop measures of implicit knowledge.

Yet another question that pertains to the question of whether there is either an

implicit – explicit model or an automatized explicit knowledge – explicit knowledge model is whether it is necessary to tease apart automatized explicit knowledge and implicit knowledge. It can be assumed that automatized explicit knowledge, if present, is functionally equivalent to implicit knowledge (DeKeyser, 2003). Irrespective of whether the hypothesised construct is implicit or automatized explicit knowledge, the type of knowledge measured by the commonly used time pressured tests (e.g. TGJTs and EI tests) was accessed with time pressure and with a low level of attention to linguistic form and the involvement of a low level of metalinguistic awareness. If one has the ability to accurately use the L2 under such conditions, it is very likely that one's L2 knowledge is at least functionally similar to implicit knowledge.

Empirically speaking, perhaps because automatized explicit knowledge and implicit knowledge are functionally equivalent, differentiating them is extremely difficult. The validation of the automatized explicit knowledge – implicit model (e.g. Suzuki, 2017) is at its early stages and more convincing empirical evidence is required. In comparison, evidence supporting the implicit-explicit model where tests with time pressure are used as measures of implicit knowledge (e.g. the EI test and the TGJT) has been reported repeatedly (R. Ellis et al., 2009; Bowles, 2011; Zhang, 2015).

Admittedly, if valid measures can be developed to measure implicit knowledge and automatized explicit knowledge, our understanding of L2 acquisition in general would be deepened. If implicit L2 knowledge that is similar to a native speaker's L1 knowledge could be developed, it would constitute evidence for the upper limit of L2 acquisition. This would call for studies that examine the nature of the knowledge measured by time pressured tests and psycholinguistic tests. Again, as indicated by current research findings, a distinction between implicit and automatized explicit knowledge is empirically difficult to prove.

To sum up, the factor analysis results which provided evidence of the construct validity of the six language measures in the present study support the hypothesised

two-factor model of implicit and explicit knowledge in accord with R. Ellis (2005), R. Ellis and Loewen (2007), Bowles (2011) and Zhang (2015).

5.2 Validation of tests of language aptitude and memory

5.2.1 The LLAMA test

In the present study, multiple measures of language aptitude and working memory were used with a view to examine how variations in aptitude and memory abilities influence the implicit and explicit knowledge of the NNS participants. In this section, the construct validity of these aptitude and memory measures is discussed first. The purpose of this part of the discussion is twofold. First, to understand the extent to which individual tests of language aptitude and working memory measure the intended constructs; and second, to examine the relationship between language aptitude and working memory.

The construct validity of the LLAMA test is supported by the results of the present study. The fact that the correlations among the four sub-tests of the LLAMA test were all statistically significant provides some evidence that what was accessed by each of the tests was correlated (see Section 4.7.1). The results of the confirmatory factor analysis (see Section 4.7.2) provide further evidence in support of the construct validity of the LLAMA test, showing that the four sub-tests were indeed measuring a unitary underlying construct, namely language aptitude.

These results contradict the findings of Granena (2012, 2013b), who suggested that LLAMA D measures a different aptitude construct from the other three sub-tests. She arrived at this conclusion based on the results of a principal component analysis, in which sub-test D loaded on a different factor to the sub-tests B, E and F in the LLAMA test. In particular, she argued that LLAMA D measures implicit learning aptitude, while the other three sub-tests measure explicit learning aptitude. However, the correlation analysis results of the present study show that this sub-test was of a

similar degree of relatedness to the other three sub-tests (r = .32, .24, and .35 with LLAMA B, E, and F respectively; $p < .05$), which means that a dissociation cannot be made between sub-test D and the other three sub-tests. The results of the factor analysis further confirmed that the LLAMA test is internally consistent, and that the four sub-tests measure different aspects of the same underlying construct. Were LLAMA D a measure of a different aptitude construct as suggested by Granena (2012), it should have loaded on a different factor. In other words, the results of the present study do not support a distinction between implicit and explicit aptitude.

However, a note of caution is due here since the error variance of the LLAMA test was quite large in the confirmatory factor analysis. Except for LLAMA B, LLAMA D, E and F had an error variance of .78, .72 and .63 respectively. This means that the hypothesised construct language aptitude could only explain a limited amount of variation in the scores of these three subtests (less than 50%), and that there were considerable measurement errors in the LLAMA battery, especially in sub-tests D and E. These measurement errors indicate that the four sub-tests of the LLAMA test were not accurate in their measurement, and thus to some extent undermine the validity of this test battery. Clearly, further validation studies of the LLAMA test are warranted.

The discrepancy between the results of the present study and that of Granena (2012; 2013b) could be attributed to the differences in the methods of analysis used. Granena (2012) conducted a principal component analysis as an exploratory factor analysis using the Varimax rotation method. As discussed earlier (see Section 5.1.3), principal component analysis does not comply with the common factor model and should not be regarded as a suitable factor analysis because it does not assume measurement error. The Varimax rotation method that Granena (2012) used did not account for factor correlation, which might also influence her results to some extent (however, she said that there was no difference in results when an orthogonal rotation method was used). In comparison, the present study conducted a confirmatory factor analysis, in which error variance was assumed and a correlated

factor relation (i.e. between language aptitude and memory) was specified. As the results of the present study suggested, there is notable measurement error (i.e. error variance) in the LLAMA test, especially in sub-tests D and E. The failure of Granena (2012) in accounting for this error variance may have influenced her results, so that a pattern of factor loading different from the present study was found.

Another aspect that may have led to a discrepancy between the results of the present study and that of Granena (2012) could be the differences in the other tests that were involved in the study. Granena (2012) used a measure of general intelligence (the GAMA test) and a measure of sequence learning ability (a Serial reaction time task), alongside the LLAMA test. The GAMA test which loaded together with LLAMA B, E and F led her to conclude that these tests were measures of explicit learning aptitude. LLAMA D loaded together with the test of sequence learning ability, which led her to conclude that these two were measures of implicit learning aptitude. The present study, however, did not include measures of general intelligence and sequence learning ability, but employed two measures of memory. Perhaps due to the different tests involved in the factor analysis, the four sub-tests of the LLAMA test loaded together in the present study while these loaded on two factors in Granena (2012).

The results indicating that the four sub-tests of the LLAMA test were testing a unitary construct, hypothesised as language aptitude, were in line with that of Rogers et al. (2017). With the purpose of validating the LLAMA test, Rogers et al., (2017) examined the relationship between individual difference factors, including age and language learning experience, and the scores of the LLAMA test. In particular, they investigated whether LLAMA D tests a different aptitude component from the other three sub-tests, following the suggestions made by Granena (2012). They hypothesised that if LLAMA D measures implicit learning aptitude, the younger learners (age 10 to 11 years old) should outperform older learners as a result of their active implicit learning system. However, Rogers et al., (2017) concluded that this hypothesis was not supported as they found that the younger learners performed

worse than the older learners in all four sub-tests. This could be an indication that what is measured by LLAMA D is similar to that measured by the other three sub-tests. In other words, the findings of Rogers et al. (2017) and the present study suggest that the LLAMA test measures a single underlying construct, that is, language aptitude.

Apart from Granena (2012) and Rogers et al. (2017), there is a lack of validation studies of the LLAMA aptitude test although this test is being utilised as a measure of language aptitude by a growing number of researchers. It is very likely that scholars use this test because it is the most recent and the only freely accessible language aptitude test. Other language aptitude tests, such as the MLAT (Carroll & Sapon, 1959) and the Hi-LAB (Doughty et al., 2010; Linck et al., 2013), are not available to individual researchers. To ensure that the construct of language aptitude is reliably measured empirically, further validation studies of the LLAMA test are in order.

5.2.2 Tests of working memory

The present study included two measures of memory, namely the NonWord (Zhao, 2015) and the SSpan (Zhao, 2015). Since only two measures of memory were included, the evidence attesting to their construct validity was limited. Some evidence can be drawn from the correlation analysis results (see Section 4.7.1), showing that the association between the NonWord test and the SSpan test was weak (r = .21, p > .05). This indicates that these two measures accessed abilities that were distinct, which is in line with the theoretical view that short-term memory and executive working memory are distinct systems (Coolidge & Wynn, 2009).

Further evidence can be seen from the results of the confirmatory factor analysis (see Section 4.7.2), in which these two tests loaded on the same hypothesised factor, that is, memory. However, the nonsignificant factor loadings of these two memory tests also suggest that the connections between them and the

hypothesised construct of memory were rather weak. One reason for this may be that the limited number of tests used to measure the memory construct influenced the estimation of factor loadings. Ideally, each hypothesised factor should have more than two indicators (i.e. three tests or more) in a factor analysis (Brown, 2015). Nonetheless, as these two measures have been widely used in testing short-term memory and working memory capacity, and both of them had an acceptable level of internal consistency as indicated by Cronbach's α estimates (see Section 3.5), it is concluded that the scores of these two tests are reliable.

5.2.3 The relationship between working memory and language aptitude

The results of the present study provide some evidence that working memory and language aptitude are distinct constructs. Firstly, the fact that there were minimal correlations between the NonWord test that measured phonological short-term memory and all four sub-tests of the LLAMA test shows that what was accessed by the short-term memory test was different from that which was accessed by the LLAMA test (see Section 4.7.1). The correlations between the working memory test and the LLAMA test were also small (r ranged from .22 to .29, $p < .05$), which again indicates that the abilities measured by the SSpan (i.e. executive working memory) and the LLAMA test were somewhat different. These findings corroborate the results of a recent meta-analysis that examines the relationship between language aptitude and working memory (Li, 2016), which also concludes that while executive working memory correlates with language aptitude, the connection between phonological short-term memory and language aptitude is very weak.

Further evidence which shows the dissociation of working memory and language aptitude is the non-significant factor correlation of .56 in the confirmatory factor analysis (see Section 4.7.2). This again suggests that the two memory tests accessed an ability that was different from that accessed by the LLAMA test. Presumably, the two memory tests tapped into a domain-general ability that is

implicated in a variety of cognitive activities (Wen & Skehan, 2011), whereas the LLAMA test accessed a domain-specific ability that is only pertinent to language learning (Li, 2016).

However, these results do not necessarily mean that working memory is not a component of language aptitude. The two hypothesised factors, namely language aptitude and memory, had a non-significant correlation of .56, which suggests that these two constructs are correlated yet relatively independent. A direct way of testing whether these two factors are components of the same hypothesised factor is to do a second-order confirmatory factor analysis. This would call for the hypothesis of another latent factor that accounts for the variance and covariance of the current two factors, that is, memory and language aptitude. However, the insufficiency of indicators for the memory factor had probably prevented the convergence of a higher order factor analysis in the present study[3]. Thus, it is concluded that the memory tests and the LLAMA test were respectively measuring two constructs, namely working memory and language aptitude.

5.3 Implicit and explicit knowledge of the NNS and the NS

The first research question asked if there is a difference between the amount of implicit and explicit knowledge of the NNS in comparison to the NS, and four hypotheses were made accordingly.

RQ1: *What is the profiling of the implicit and explicit knowledge of the NNS and the NS? Is there a difference between the implicit and explicit knowledge of the NNS in comparison to the NS?*

Hypothesis 1a: The NNS will have more explicit knowledge than implicit knowledge in English.

Hypothesis 1b: The NS will have more implicit knowledge than explicit

[3] A higher order confirmatory factor analysis of the model of language aptitude and memory was attempted but this higher model did not converge.

knowledge in English.

Hypothesis 1c: The level of implicit knowledge of the NNS will be lower than that of the NS.

Hypothesis 1d: The level of explicit knowledge of the NNS will be higher than that of the NS.

Hypothesis 1a is supported by the results of the present study. The NNS participants indeed had more explicit than implicit knowledge, which is shown by the higher weighted sum scores of explicit knowledge than implicit knowledge in the form of percentages (.63 of implicit knowledge in comparison to .74 of explicit knowledge; see Section 4.6 and Table 4.41). The results of the paired samples *t*-tests further revealed a significant difference between the implicit and explicit knowledge of the NNS ($t(84) = -8.47$, $p < .01$). The associated large effect size (Cohen's $d = .92$) of the results of the paired samples *t*-test suggests that the difference between implicit and explicit knowledge of the NNS was substantial.

Hypothesis 1b which predicted that the NS will have more implicit knowledge than explicit knowledge is also supported. Judging from the weighted sum scores of implicit and explicit knowledge in the form of percentages of the NS (.89 of implicit knowledge and .81 of explicit knowledge, see Section 4.6 and Table 4.41), it can be seen that the NS indeed had a higher level of implicit knowledge than explicit knowledge. The results of the paired samples *t*-test provided further evidence of a statistical difference between the implicit and explicit knowledge of the NS ($t(18) = -8.47$, $p < .01$). These findings were expected since the knowledge of the NS was largely implicit. At the time of data collection, the researcher asked the NS if they had been taught English grammar at any time, and most of the NS participants claimed that they only had some experience in learning word classes and the basics of sentence structure. In terms of grammar rules, they did not have the experience of being taught explicitly. In other words, the NS participants had not had many opportunities to develop explicit knowledge. Therefore, a significant difference

between implicit and explicit knowledge was observed.

Hypothesis 1c made the prediction that the NS will have more implicit knowledge than the NNS. This prediction is supported by the results of the independent samples *t*-test, showing that difference in the level of implicit knowledge between the NS and the NNS is statistically significant ($t(93.90) = 16.16$, $p < .01$; see Section 4.6). This between group difference in implicit knowledge is substantial, as indicated by the large effect size; Cohen's $d = 2.82$. These findings suggest that despite having been immersed in the L2 environment for a rather long period of time (LoR > 8 years), the NNS participants as a group still lacked implicit knowledge in comparison with the NS group.

Hypothesis 1d made the prediction that the NNS will have more explicit knowledge than the NS. At first glance, it might seem that this hypothesis is not supported because the NS obtained a higher weighted mean score of .81 in explicit knowledge than the NNS (.74) (see Section 4.6 and Table 4.41). The paired samples *t*-test results further showed that there was a significant between-group difference in explicit knowledge (t (102) = 2.79, p = .006). However, these results could be attributed to the fact that this explicit knowledge score was comprised of the scores of three different tests, namely the UGJT, the MKT and the cloze test. As discussed earlier (see Section 5.1.2), it was possible for the participants, and especially the NS participants, to use implicit knowledge in the completion of the UGJT and the cloze test. Therefore, it can be argued that the weighted explicit knowledge score may not fairly represent the explicit knowledge of the NS. In other words, it is inappropriate to conclude that the NS had a higher level of explicit knowledge than the NNS based on the explicit knowledge scores which included the scores of the UGJT and the cloze test, since both of these tests could be completed with the use of implicit knowledge.

Alternatively, perhaps a better way of reaching valid conclusions about the level of explicit knowledge of the NS is to only consider their scores of the MKT. Unlike the UGJT and the cloze test, the MKT is a purer measure of explicit knowledge since

implicit knowledge is unlikely to be employed in this test. The results of the independent samples *t*-test using the scores of the MKT showed that the NNS group did perform better than the NS group (see Section 4.2 and Table 4.5). This finding lends support to hypothesis 1d that the NNS had more explicit knowledge than the NS.

The statistically lower level of implicit knowledge of the NNS in comparison to the NS highlights how difficult the development of implicit knowledge can be, especially considering the fact that the NNS in the present study had been immersed in an English-speaking environment for a rather long period of time. This finding is probably an indication that what can be acquired implicitly by L2 learners from communicative contexts is quite limited (N. C. Ellis, 2008). The average LoR of the NNS was 15.32 years (see Table 4.63), which means that the NNS had the opportunities of being exposed to ample linguistic input. Moreover, as reported by the NNS (n = 86), 77 of them had the experience of studying in New Zealand universities for an average of 3.47 years, and some others even also had the experience of receiving primary (n = 20) and secondary (n = 34) education in New Zealand (see Table 4.63). In addition, the NNS reported an average of 7.64 years of working in New Zealand where English was used as the working language. These self-reported data seem to imply that there were chances of developing a relatively high level of implicit knowledge during years of immersion in the L2 context for the NNS, yet the overall implicit knowledge scores indicated otherwise. The investigation of the language use situations of the NNS appears to have provided some insight for why this was the case, as it can be seen that many participants (n = 72) speak more Chinese than English with their family (see Table 4.66). It is possible that the language being used at work was profession-specific, and that the NNS did not engage in social activities to increase their overall linguistic competence as most of them mainly socialize with other Chinese people (n = 64). Although a lot of participants also claimed that they socialize with their English-speaking friends regularly (n = 56) and that they like to attend local activities and to talk to Kiwis in

English (n = 49), their implicit knowledge fell short when being tested on a series of demanding linguistic tests (i.e. the EI test and the TGJT).

In summary, the analysed results show that the profiling of implicit and explicit knowledge of the NS and the NNS was as predicted. The NNS had more explicit knowledge than implicit knowledge, whereas the NS had more implicit knowledge than explicit knowledge. When these two groups were compared, the NNS had much less implicit knowledge than the NS, but they had more explicit knowledge than the NS.

5.4 Age, length of residence and L2 knowledge

The second research question asked about the relationship between learning age and implicit and explicit L2 knowledge. Two different age variables, namely AO and AoA, were investigated. AO refers to the age when the learning of English begins, and AoA refers to the age when immersion in the L2 context begins. In the case of the NNS in the present study, AO generally refers to the age when they started learning English in China, whereas AoA refers to the age when they immigrated to New Zealand. As a factor relevant to age, the role of LoR, which refers to the length of residence in the L2 country, was also examined. The research question and hypotheses are:

RQ2: *What is the relationship between age (AO, AoA), LoR and implicit and explicit L2 knowledge?*

Hypothesis 2a: AO will not influence the implicit knowledge of the NNS.

Hypothesis 2b: AoA will have a negative influence on the implicit knowledge of the NNS.

Hypothesis 2c: LoR will have a positive influence on the implicit knowledge of the NNS.

Hypothesis 2d: AO will not influence the explicit knowledge of the NNS.

Hypothesis 2e: AoA will not influence the explicit knowledge of the NNS.

Hypothesis 2f: LoR will not influence the explicit knowledge of the NNS.

Hypothesis 2a made the prediction that AO will not influence the implicit knowledge of the NNS. This hypothesis is supported by the results of the multiple regression analysis, which clearly demonstrated that AO (β = -.24, p = .07; see Table 4.50) was not a significant predictor of implicit knowledge (see Section 4.9.2). However, a negative relationship between AO and implicit knowledge (r = -.48, p < .01; see Table 4.49) was found in the correlation analysis, which was some indication that there was a connection between AO and implicit knowledge. To reconcile these seemingly contradictory results, the results of regression analysis were considered more robust because it revealed a causal relationship between the independent variable AO and the dependent variable implicit knowledge. In comparison, although the results of correlation analysis indicate a significant correlation between the two variables, it does not necessarily imply causation. In other words, multiple regression analysis provides more direct results in relation to the extent to which AO predicts implicit knowledge of the NNS, and it can be seen that the starting age of learning English (i.e. AO) did not predict long-term L2 implicit knowledge.

Similarly, the results of the multiple regression analysis (see Section 4.9.2) lend support to hypothesis 2b, which predicted that AoA will influence the implicit knowledge of the NNS. Evidently, AoA appeared as a significant predictor of implicit knowledge (β = -.36, p = .006, see Table 4.50). This means that age of immersion in the L2 context is a factor that can account for some variance in long-term L2 implicit knowledge, and that there is a negative causal relationship between age of immersion and long-term implicit knowledge. This negative relationship was also evident in the results of the correlation analysis, as is shown by the significant correlation coefficient between AoA and implicit knowledge (r = -.53, p < .001; see Table 4.49).

In contrast, hypothesis 2c which predicted that LoR will influence the implicit knowledge of the NNS is not supported. The results of the correlation analysis showed that there was no correlation between LoR and implicit knowledge (r = .08, p > .05, see Section 4.9.1 and Table 4.49). In addition, the results of the multiple regression analysis provided further evidence that LoR was not a significant predictor of implicit knowledge (β = .11, p = .24, see Table 4.50). Therefore, it was concluded that hypothesis 2c is not borne out in the present study.

Hypothesis 2d predicted that AO will not influence the explicit knowledge of the NNS, and this hypothesis is supported by the analysed data. Both the results of the correlation analysis and the regression analysis provided some evidence, that AO barely had any correlations with explicit knowledge (r = .05, p > .05; see Table 4.49) and that AO was not a significant predictor of explicit knowledge (β = -.06, p = .68; see Table 4.51). These results show that initial learning age of English in an instructed context does not influence the amount of explicit knowledge developed in the long run (Huang, 2016).

Similarly, the present results lend support to hypothesis 2e, which predicted that AoA will not influence the explicit knowledge of the NNS. Both the results of the correlation and the regression analysis provided some evidence in this respect, showing a very weak correlation between AoA and explicit knowledge (r = .09, p > .05; see Table 4.49) and that AoA was not a significant predictor of explicit knowledge (β = .14, p = .68; see Table 4.51). These results demonstrate that like AO, AoA also did not have much association with long-term explicit knowledge.

The last hypothesis, 2f, addressed the relationship between LoR and explicit knowledge. It was predicted that LoR will not influence explicit knowledge, and the current results lend support to this hypothesis. The correlation analysis results show that there was a weak and nonsignificant correlation between LoR and explicit knowledge (r = .16, p > .05; see Table 4.49). In addition, the regression analysis results suggested a lack of predictive power of LoR on explicit knowledge, as indicated by the nonsignificant regression coefficient β = .17, p = .13 (see Table

4.51). Based on these results, it was concluded that LoR was not related to explicit knowledge.

In summary, the present study provided some evidence that regarding the relationship between AO, AoA, LoR and implicit L2 knowledge, AoA was the only significant predictor of implicit knowledge. AO and LoR, in comparison, could not predict long-term implicit knowledge. In relation to explicit knowledge of the NNS, none of these three variables were significant predictors. These findings corroborate those of many previous studies, which repeatedly show that age of immersion in the L2 context (i.e. AoA) is the strongest predictor of long-term L2 learning outcome in comparison to initial learning age (i.e. AO) and LoR (see Birdsong, 2006 for review).

One possible way of understanding the different roles of AoA and AO is from the perspective of the variation in linguistic input in different language learning contexts. The majority of the NNS (n = 75) participants had the experience of learning English both in an instructed context (i.e. China) and a naturalistic context (i.e. New Zealand). They were exposed to English at a rather young age (mean AO = 10.15, see Table 4.63) in the instructed context. However, this young starting age of learning English did not guarantee the development of a high level of implicit knowledge. Instead, it was the age at which they were immersed in an English-speaking environment, which was much older on average (mean AoA = 21.40, see Table 4.63), that predicted long-term achievement in implicit knowledge. Presumably, the amount and quality of linguistic input that the NNS received in China differed from that which they received after immigration. Differences in the amount and quality of linguistic input also imply that there were differences in the activities that the NNS participated in using English in the instructed context in comparison with the naturalistic context. As reported by the NNS in the questionnaire, the learning activities that they engaged in while in China were based on textbooks, rather than communication. This means that the NNS might have received instruction which was focused on explicit knowledge and understanding of

English grammar, while their opportunities of using English to communicate were minimal (see Table 4.64). This lack of ability to use English communicatively is also shown in their self-reported language proficiency after immigration (see Table 4.65), as many participants reported that they encountered difficulties in speaking, reading and writing in English. In contrast, after being immersed in the naturalistic context, their opportunities for communicating in English increased as the majority attended New Zealand universities (n = 77) and also worked using English as the working language (n = 83). Therefore, the significant role of AoA implies that the younger the age at which one was immersed in a naturalistic learning context, the earlier one started to receive massive quality linguistic input and to use English in communication, with the result that more implicit knowledge was developed.

However, if differences in the amount and quality of input were the cause of the different roles played by AO and AoA, then one would expect LoR to be a significant predictor of implicit knowledge as well. The logic behind this is simply that a longer LoR would denote a larger amount of input. Yet, as is shown earlier, LoR was not a predictor of implicit knowledge. This lack of correlation between LoR and implicit knowledge could be a result of the requirement of more than 8 years of LoR in the present study, which prevented it from appearing as a significant predictor. Had some NNS with shorter LoR been sampled, different results might be obtained. However, as the primary focus of the present study is to examine the long-term learning L2 achievement of the NNS, the requirement of more than 8 years of LoR is regarded as essential. Similar findings have been reported by several previous researchers (DeKeyser, 2000; Flege et al., 1995; 1999; Larson-Hall, 2001; Thompson, 1991), who also found that LoR could not predict L2 learning outcomes. In studies that controlled for a minimum LoR (e.g. > 5 years or 10 years) for the purpose of examining long-term learning outcomes, it has generally been suggested that LoR does not influence L2 knowledge in the long run. This lack of relationship between LoR and implicit L2 knowledge shows that the amount and quality of input cannot account for variation in NNS's implicit knowledge, which in turn highlights

the significant role of age, the maturational factor.

So far in the discussion, it has been argued that the maturational factor age profoundly influences L2 learning outcomes. On the other hand, age of immersion seems to have a more salient role for implicit knowledge than age of learning, and so the researcher came to wonder how the different roles of these two age variables can be explained. As the possibility of the influence of input has been excluded, the other explanation left to be considered concerns the different types of learning involved (i.e. implicit or explicit learning) in association with AoA and AO. In the case of the NNS participants in the present study, when they learnt English in the instructed context, they mainly relied on their explicit learning system because they were instructed explicitly in classroom settings. It can be seen from the self-reported learning experiences that English was taught in a deductive way so that grammar rules and vocabulary were made explicit for the student to remember (see Table 4.64). Similarly, the NNS learnt English speaking and listening by reading after the teacher and doing text-book based exercises. This kind of explicit learning would very likely lead to the development of explicit knowledge. In comparison, the NNS might have mainly engaged in implicit learning in the naturalistic context because they studied and worked in New Zealand, which implies that they were learning English by using it (see Table 4.65). Some NNS participants (n = 26) attended language schools which were unlikely to be teaching English explicitly with a heavy focus on grammar. Indeed, the majority of these participants (n = 24) reported that in language schools they learnt how to use English to communicate rather than learning grammar. Based on these self-reported data, it can be assumed that the NNS were more likely learning implicitly after being immersed in the naturalistic context.

Therefore, the different roles of AO and AoA could be a manifestation of the results of explicit and implicit learning respectively, with AO associated with explicit learning in an instructed context and AoA linked to implicit learning in a naturalistic context. Further evidence that lends support to this explanation comes from the fact that AoA was only found to be influential for implicit knowledge, not for explicit

knowledge (see Section 4.9.2). Specifically, the negative linear relationship between AoA and implicit knowledge shows that the earlier the AoA, the more the implicit knowledge developed. As presented earlier, the NNS mainly relied on implicit learning after being immersed in the L2 context, so it can be concluded that further development of implicit knowledge was primarily a result of implicit learning. In other words, this finding provided some empirical evidence that age only influences the implicit learning system (DeKeyser, 2003; Lenneberg, 1967), in the sense that one's ability in learning through mere exposure and communicative interaction declines as age increases. As the implicit learning system becomes less active with the increase of age, there would be a decrease in the level of the primary outcome of implicit learning process, namely implicit knowledge. As is shown by the present study, a younger AoA implies better access to the implicit learning system, which in turn leads to a higher level of implicit knowledge developed after a long length of residence in the L2 country. The lack of relationship between AoA and explicit knowledge further shows that this age effect is restricted to the implicit learning system.

The negative influence of age on implicit knowledge is no surprise since there has been a wealth of research findings that suggest age is negatively correlated with L2 learning outcomes (e.g. Abrahamsson & Hyltenstam, 2008; 2009; Abrahamsson, 2012; DeKeyser, 2000; DeKeyser et al., 2010; Granena & Long, 2013a; Hyltenstam & Abrahamsson, 2000; Johnson & Newport, 1989; Long, 1990). However, this negative linear relationship between AoA and implicit knowledge provides some counter evidence for the hypothesised critical period or sensitive period. The shape of the age influence was noticeably linear, as indicated by the negative regression coefficients when age was regarded as both a continuous variable (see Figure 4.11) and a categorical variable (see Figure 4.12). In other words, the present study did not find a finite period of time, during which second language acquisition is advantageous as suggested by the critical/sensitive period. The present study also did not find evidence of a clear discontinuity, showing that there is qualitative difference

in terms of the final outcomes of L2 acquisition (Birdsong, 2006; Granena & Long, 2013a). Therefore, it can be concluded that the influence of age on implicit knowledge is better viewed as general maturational constraints.

In summary, it is found that age of immersion has an influence on implicit knowledge, whereas the age that one begins to learn English in a foreign language context does not have an influence on implicit knowledge. The significant influence of age of immersion is very likely a result of the type of learning taking place, because an early age of immersion means better access to the implicit learning system. This access to the implicit learning system overrides the influence of length of residence, as indicated by the lack of relationship between LoR and implicit knowledge.

5.5 Aptitude, working memory and L2 knowledge

The third research question asked what the relationship is between language aptitude, working memory and L2 knowledge. Four hypotheses were made in light of previous research.

RQ 3: *What is the relationship between language aptitude, working memory and implicit and explicit L2 knowledge?*

Hypothesis 3a: Language aptitude will have a positive influence on the implicit knowledge of the NNS. This influence is mainly contributed by sequence learning ability as measured by LLAMA D.

Hypothesis 3b: Language aptitude will also have a positive influence on the explicit knowledge of the NNS. This influence is contributed by language analytic ability as measured by a combination of LLAMA B, E and F.

Hypothesis 3c: Working memory will have a positive influence on both the implicit knowledge and the explicit knowledge of the NNS.

Hypothesis 3a and 3b made predictions regarding the influence of language aptitude on implicit and explicit knowledge respectively. As part of the hypotheses, the specific roles of aptitude components (i.e. sequence learning ability and analytic ability) were predicted in light of previous research (Granena, 2012). However, as the present study failed to dissociate language aptitude into different components (see Sections 4.7 and 5.2), it was impossible to investigate the relationships between particular aptitude components and L2 learning outcomes. Instead, by viewing language aptitude as a holistic construct (i.e. by calculating a factorial sum score of the LLAMA test, see Table 4.45), its relationships with implicit and explicit knowledge (see Section 4.8.1) were examined by doing a correlation analysis. The predictive power of language aptitude on implicit and explicit knowledge was further examined by conducting two regression analyses (see Section 4.8.2).

Hypothesis 3a made the prediction that language aptitude will influence the implicit knowledge of the NNS, and this hypothesis is confirmed. However, since the language aptitude scores were the weighted sum scores of four sub-tests, it can be argued that the influence of language aptitude on implicit knowledge was contributed by various aptitude components. The results of the correlation analysis provided some evidence in this respect, as it can be seen that there was a significant correlation between language aptitude and implicit knowledge ($r = .55$, $p < .01$; see Table 4.46). Several significant correlations between the sub-tests of LLAMA and tests of implicit knowledge (i.e. the EI test and the TGJT) can also be seen (see Section 4.8.1 and Table 4.44). The results of the regression analysis provided further evidence that language aptitude has an influence on implicit knowledge ($\beta = .57$, $p < .001$; see Table 4.47).

The results of the present study also confirmed hypothesis 3b, which predicted that language aptitude will have an influence on the explicit knowledge of the NNS. Evidence in support of this comes from the results of both the correlation analysis and the regression analysis. There were significant correlations between language aptitude and explicit knowledge ($r = .34$, $p < .01$; see Table 4.46), and between

certain LLAMA sub-tests and tests of explicit knowledge (see Table 4.44). The results of regression analysis further showed that language aptitude was a significant predictor of explicit knowledge (β = .39, p = .001; see Table 4.48). This evidence means that language aptitude, as represented by a combination of several cognitive abilities, including sound recognition, sound-symbol pairing and grammatical inferencing ability, has an effect on explicit knowledge. This conclusion is in line with the considerable research literature that shows the implication of language aptitude in explicit and conscious learning conditions (See Li, 2016 for a recent meta-analysis review). Therefore, it is not surprising to see a correlation between language aptitude and the product of conscious learning, that is, explicit knowledge.

The fact that there were significant correlations between language aptitude and L2 knowledge implies that the influence of language aptitude is not only evident in how fast a learner can learn a foreign language as previous research has shown (Carroll & Sapon, 1959), but also in how well the learner learns the L2. The fact that language aptitude was a significant predictor for both implicit and explicit knowledge provides further evidence that language aptitude is related to the outcomes of second language acquisition. In other words, the findings of the present research emphasise the centrality of language aptitude to L2 learning (de Graaff, 1997; Robinson, 1995a), in the sense that it is not only related to the speed of language acquisition, but also to success in acquisition.

One important finding of the present study is that language aptitude appeared to have a stronger association with implicit knowledge than with explicit knowledge. Evidence in support of these findings comes from the higher correlation between aptitude and implicit knowledge (r = .55, p < .001) than between aptitude and explicit knowledge (r = .34, p < .001) (see Table 4.46). Further evidence from the results of the regression analyses also show that the predictive power of language aptitude for implicit knowledge (β = .57, p < .001) is higher than that for explicit knowledge (β = .39, p < .001) (see Tables 4.47 and 4.48). This stronger association between language aptitude and implicit knowledge is a different finding from

previous research as it has been suggested by a recent meta-analysis that language aptitude is an explicit construct that relates to explicit learning (Li, 2015). This is probably a result of methodological discrepancies in the way that learning outcomes are measured, as previous research did not make a distinction between implicit and explicit knowledge. As discussed earlier (see Section 5.1), variations in test requirements (e.g. a timed vs. untimed test) could lead to the use of distinctive types of L2 knowledge. For example, a TGJT would primarily lead to the access of implicit knowledge, while an UGJT is more likely to require the use of explicit knowledge. In fact, in similar studies conducted in naturalistic learning contexts (e.g. DeKeyser, 2000; DeKeyser et al., 2010), both learning outcomes and language aptitude were measured with a bias towards the use of explicit processes, with the former mainly accessing explicit knowledge and the latter requiring explicit analytical abilities. It is probably this limitation in accessing implicit knowledge that has led to the conclusion that language aptitude is a largely explicit construct.

The investigation of the role of language aptitude on NNS's implicit knowledge is particularly relevant to the understanding of the role of language aptitude in naturalistic contexts. Although it cannot be assumed that the implicit knowledge of the NNS in the present study was developed entirely through immersion in the L2 country, it is possible that this was the case or that at least some implicit knowledge was developed. The significant correlation between language aptitude and implicit knowledge (r = .55, p < .01; see Table 4.46) and the significant standardised regression coefficient ((β = .57, p < .001; see Table 4.47) therefore, provide evidence that there is a chance for aptitude to facilitate the development of implicit knowledge, probably through its contribution in naturalistic learning conditions. Although a formidable body of evidence suggests that individual differences in language aptitude are more relevant to explicit learning processes (see Li, 2015; 2016; for meta analyses), the high correlation between language aptitude and implicit knowledge found in the present study demonstrates that better language aptitude would lead to a higher level of implicit knowledge. As reported by the NNS (see

Table 4.63), it was very likely that they mainly engaged in implicit learning while being extensively immersed in the L2 context (LoR > 8 years). Under such a premise, the fact that language aptitude significantly predicts implicit L2 knowledge provides some evidence that even under implicit learning conditions, language aptitude plays a role. This finding corroborates the results of several previous experimental studies, in which language aptitude is found to be relevant to both implicit and explicit learning conditions (de Graaff, 1997; Robinson, 1995a). As argued by Skehan (1986; 1989; 1998; 2002), language aptitude should play an equally important role in more implicit and naturalistic learning contexts because it reflects a general ability to process linguistic input. In naturalistic or implicit learning conditions with ample input, greater language aptitude would bring an advantage in the processing of input and thus lead to better learning gains.

Hypothesis 3c predicted that working memory will influence both the implicit and the explicit knowledge of the NNS, but this hypothesis is not supported by the present study. Working memory was operationalised as a combination of two memory abilities, namely executive working memory as accessed by the SSpan, and short-term memory as measured by the NonWord (see Sections 3.4.4). Following the results of the confirmatory factor analysis (see Section 4.7.2), a weighted composite score of working memory was computed (see Table 4.45). The relationship between working memory and L2 knowledge was then examined by conducting correlation and regression analyses (see Section 4.8). The analysed results suggested that working memory was not related to implicit knowledge, as there were very weak correlations between the two (r = .11, p > .05, see Tables 4.44 and 4.46). This lack of association between working memory and implicit knowledge was later confirmed in the regression analysis, as working memory was not found as a significant predictor of implicit knowledge (β = -.08, p = .40, see Table 4.47). Similarly, both the results of the correlation analysis and the regression analysis suggested a lack of association between working memory abilities and explicit knowledge. Working memory barely had any correlations with explicit knowledge (r = .03, p > .05; see Tables 4.44 and

4.46), and it was not a significant predictor of explicit knowledge (β = -.10, p = .36, see Table 4.48). These results indicate that variations in working memory ability do not lead to variations in explicit knowledge.

The lack of correlation between working memory and L2 knowledge could be due to the design of the study. Much of the previous research that supports the role of working memory adopted an experimental, or a "macro" approach (Skehan, 2016), by administering some working memory test, collecting achievement scores after treatment at later stages and then exploring the correlations over the time period. In this approach, some sort of comparison was generally made between the performances of the participants over a learning period. If an improvement is observed, working memory is then examined for its contribution to this improvement (e.g. Cheung, 1996; French, 2006). These studies have in common that the achievement scores represent the rate of acquisition, rather than the outcomes. Therefore, findings obtained by these studies are more pertinent to the evaluation of working memory's role in short-term language learning outcomes, and any generalisation of these findings to the predictive power of working memory in long-term L2 acquisition should be made with caution.

The present study is a cross-sectional study that distinguishes itself from the previous studies. In the present study, the outcomes of L2 acquisition, rather than the rate of acquisition, were examined by a series of tests of implicit and explicit L2 knowledge. Together with the examination of L2 learning outcomes, two different working memory components, namely phonological short-term memory and executive working memory, were assessed in order to examine the role of memory in long-term L2 acquisition. The fact that there was no correlation between working memory ability and explicit L2 knowledge indicates that neither of these two components appears to influence how well an L2 can be developed in the long run. This finding indicates that the contribution of working memory to long-term L2 acquisition achievements seems to be very limited.

The lack of correlation between working memory and L2 knowledge could also

be due to the fact that the NNS in the present study were relatively proficient, albeit far from native-like, in English. Previous research has shown that the effects of working memory are sensitive to learners' proficiency and L2 acquisition stages (Gass & Lee, 2011; Juffs & Harrington, 2011; Wen, 2016), and that as learners' level of proficiency increases, the facilitative role of working memory decreases. In a longitudinal study that includes measures of L2 gains at different time points, the strongest effect of working memory was found within the first 3.5 months of learning, after which this impact gradually lessened with the increase of learners' proficiency (Serafini & Sanz, 2016). These findings suggest that the influence of working memory is most likely to be found at the early stages of L2 acquisition when learners' proficiency is low. The NNS in the present study had well passed the initial stage of L2 acquisition at the time of participation because they had been living in the L2 country for more than 8 years and the majority of them had received several years of English instruction in China (see Table 4.63). The NNS also reported that they had received an average of 3.47 years of university education and had been using English for work for an average of 7.64 years in New Zealand (see Table 4.63). Based on this data, it could be assumed that the NNS were probably quite proficient in their L2 and that they had at least an upper intermediate level of English proficiency, without which studying in a New Zealand university or working in English would be difficult. The lack of correlation between working memory and L2 knowledge therefore, echoes previous research findings that show that working memory no longer appears as a significant predictor at later stages of L2 acquisition where L2 proficiency is relatively high.

The different roles that language aptitude and working memory play on implicit and explicit knowledge, as demonstrated by the results in this study, could be attributed to the assumption that language aptitude is domain specific while working memory is domain general. Language aptitude is domain specific in the sense that it relates only to language learning (Li, 2016). In comparison, working memory capacity is domain general since it is one of the general learning mechanisms (Wen

& Skehan, 2011), which is implicated in various activities other than language learning. This distinction could be another source for the different contributions made by language aptitude and working memory to implicit and explicit knowledge. Differences in domain specific abilities, that is, language aptitude, are specifically relevant to the learning of an L2, therefore bring about differences in the knowledge of the L2. Working memory, whilst involved in the L2 learning process, is perhaps particularly relevant to the early stages of L2 learning as a domain general ability and thus its relation to long-term L2 acquisition outcomes is less noticeable.

Given the results of the present study, it is concluded that language aptitude has an influence on both implicit and explicit knowledge, whereas working memory abilities do not influence long-term L2 learning outcomes. It is very likely that language aptitude not only plays a role for explicit learning, but also for implicit learning.

5.6 Age, aptitude and L2 knowledge

The final research question asked what the relationship is between age, aptitude, and long-term learning outcomes. Particularly, this research question attempted to examine if there is an interaction between age and language aptitude.

RQ4: *What is the joint influence of age and language aptitude on implicit and explicit L2 knowledge? Is there an interaction between age and aptitude?*

Hypothesis 4a: Both age and language aptitude will have an influence on the implicit knowledge of the NNS.

Hypothesis 4b: Language aptitude will influence the explicit knowledge of the NNS, but age will not.

Hypothesis 4c: There is an interaction between age and language aptitude. Age will negatively influence the implicit knowledge of the NNS, but this influence will be mitigated by high language aptitude.

Hypothesis 4a predicted that both age and language aptitude will have an influence on the implicit knowledge of the NNS and this hypothesis is confirmed. These two factors jointly explained 42% of variance in implicit knowledge (R^2 = .42, see Section 4.10.1), and both were significant predictors of implicit knowledge (see Table 4.54). While age of immersion had a negative influence on implicit knowledge (β = -.38, p < .001), language aptitude had a positive influence (β = .39, p < .001). However, the magnitudes of the influences of age and language aptitude were roughly the same, judging from the partial R^2 statistics (partial R^2 = .17 and .18 of AoA and language aptitude, respectively; see Table 4.54). Similar results were suggested when age and language aptitude were examined as categorical variables, with increases in AoA groups leading to significant decreases in implicit knowledge and increases in language aptitude leading to increases in implicit knowledge (see Table 4.56). However, when treated as categorical variables, age appeared to have had a stronger influence on implicit knowledge than language aptitude, judging from the larger partial η^2 of .17 for age and .11 for language aptitude respectively.

The present study partly supports hypothesis 4b, which predicted that language aptitude will influence the explicit knowledge of the NNS, but age will not. It was found that both AoA and language aptitude were significant predictors of explicit knowledge, as is shown by the significant standardised regression coefficients of -.28 of AoA and -.49 of language aptitude (p = .011 and p < .001, respectively, see Table 4.55). Overall, these two factors explained 20% of variance in explicit knowledge. The direction of the influence of these two factors on explicit knowledge was the same, both of which had a positive influence. However, language aptitude had a much larger influence in magnitude than age, as is shown by the partial R^2 statistics of .20 of aptitude in comparison to .08 of age. When treated as a categorical variable, it was found that the significant influence of age was likely a result of the participants in the late group (AoA > 30) because only this group differed significantly from the early group (AoA < 13) in explicit knowledge (see Table 4.57).

In other words, the statistically significant regression coefficient of age may be an artefact of the very good performance of the NNS in the late group. In comparison, significant increases in explicit knowledge were observed as the level of language aptitude increased (see Table 4.57). Again, language aptitude had a much larger influence on explicit knowledge than age, as indicated by the larger partial η^2 of .21 in comparison to .11 of age.

Hypothesis 4c predicted that there will be an interaction between age and language aptitude, but his hypothesis is not borne out. Specifically, the present study attempted to examine if a high level of language aptitude would mitigate the negative influence of age, but the results failed to find evidence for this. When a model with the main effects of age and aptitude and the interaction effect between age and aptitude (i.e. model AAM3) was compared to the regression model with the main effects of age and aptitude (i.e. model AAM2), the predictive power of the model with the interaction effect was lower than the model without the interaction (see Table 4.58). These results provided evidence that while age and language aptitude contribute to L2 knowledge independently, there was not an interaction between the two. However, it is possible that the failure in finding a significant interaction effect between age and aptitude was because of the lack of participants with a high level of aptitude. As can be seen from Table 4.59, teen and adult learners (AoA > 13 and 20, respectively) with an outstanding level of language aptitude did not exist in the data, yet these learners were essential to determine whether immersion in a naturalistic context at a later age could be compensated for with high language aptitude. However, it seems that such learners were difficult to find using the sampling strategy in the present study. Thus, based on these results, it was concluded that hypothesis 4c is not supported.

The finding that both age and language aptitude had an influence on the implicit knowledge of the NNS has two implications. First, for people with the same age of immersion, higher levels of language aptitude would be an advantage. This means that even for learners who have an early age of immersion (e.g. before puberty),

higher levels of language aptitude would probably lead to the development of a higher level of implicit knowledge. This finding is in line with the findings of Abrahamsson & Hyltenstam (2008) and Granena (2014), who also found that language aptitude plays a role for early learners. In turn, the present finding is in keeping with the assumption that "language aptitude is a significantly advantageous condition for attaining a level of L2 proficiency identical to that of native speakers for child learners" (Abrahamsson & Hyltenstam, 2008, p. 503). Some direct evidence is provided by the present study because implicit knowledge was separately measured and there were 9 participants whose implicit knowledge scores fell in the range of the NS (see Section 4.11). Of these 9 participants, 8 had an AoA before the age of 13 (i.e. puberty). The only exception was participant 4, who had an AoA of 18. These findings demonstrated that learners who had been immersed in the L2 context at an early age indeed had a chance to become near-native, whereas those who were immersed after puberty generally failed to do so despite a long length of residence in the L2 country. However, it should be noted that language aptitude must have been a determining factor for these learners because not all who were immersed before 13 years old achieved near-native levels in implicit knowledge (see Table 4.61). In fact, of the 23 participants in the young group, 15 of them failed to become near-native (see Table 4.62). Table 4.62 showed that many of these participants (n = 8) had average language aptitude, and the four participants who were close to the lowest implicit knowledge score of the NS group had a good or an excellent level of language aptitude. Moreover, within the near-native NNS participants, the majority of them (n = 6, see Table 4.61) also had an above average level of language aptitude. These findings imply that even for young learners, those who have better language aptitude would be at an advantage in developing implicit L2 knowledge.

Second, for learners with the same level of language aptitude, an early age of immersion would be advantageous for the development of implicit knowledge. This is probably because implicit knowledge is largely a result of implicit learning, and learners who are immersed at a young age may have better access to their implicit

learning system (see Section 5.4). As learners' age of immersion increases, the robustness of the implicit learning system gradually decreases, which would lead to less implicit knowledge being developed. In this process, a higher level of language aptitude could facilitate, but could not guarantee, the development of a high level of implicit knowledge. However, the facilitative role of language aptitude might be as influential as age, as indicated by the similar standardised regression coefficients (β = -.38 and .39 of age and aptitude, respectively, see Table 4.54) and the partial R^2 statistics (R^2 = .17 and .18 of age and aptitude, respectively, see Table 4.54).

In the present study, participant 4 appeared as someone who had overcome the negative influence of a late immersion age (AoA = 18, see Table 4.61), and reached a NS level of implicit knowledge. This participant reported that she started learning English at the age of 7 in China and had received ten years of classroom instruction before moving to New Zealand. She also reported that she benefited from attending extra-curriculum English classes where she spent four hours for homework and reading per week for three years from the age of 13 to 16. On top of this, she spent one to two hours per week with a foreign teacher (i.e. an English teacher who was a native speaker of English) and listened to CNN to practise English speaking and listening. She considered that these learning experiences were crucial in enabling her to communicate with Kiwis without major difficulties. It can be seen that before immersion, this participant had already put a tremendous amount of effort into learning English. This effort probably exceeded that of the other participants who only learnt English at school. After immersion, she spent eight years altogether in a New Zealand university, where she obtained her bachelor's and master's degrees. At the time of data collection, this participant was close to the completion of her doctoral degree, and at the time of the writing of this dissertation, she has graduated with her doctoral degree. During these years of receiving education in New Zealand, it could be assumed that this participant continued to develop her implicit knowledge, with the help of her good level of aptitude, and mainly learnt English implicitly. Such implicit learning very likely also took place in her daily life, as she reported

that she socialises with her English-speaking friends regularly, that she speaks more English than Mandarin with her flatmate in New Zealand, and that she often watches English TV shows and movies. At the same time, she did not indicate on the questionnaire that she engaged in activities in which she would use more Mandarin than English. This means that this participant probably learnt a lot through using English on a daily basis, so that she was able to reach a native speaker level of implicit knowledge after fourteen years of immersion.

Nonetheless, this single case of reaching near native level of implicit knowledge by no means undermines the profound negative influence of age. The fact that only one participant out of 86 participants with an age of immersion older than puberty was found to be near native provides some evidence that nativelikeness in late second language acquisition is not typical (Marinova-Todd, 2003). As presented in the previous paragraph, it can be seen that this participant put in much effort in learning English before immersion, and probably continued to do so after immersion. In addition, her experience of had had the chance to learn more implicitly, including watching CNN and communicating with a native English speaker, had probably brought about positive influence on her learning outcomes. She must have also benefited from her high level of language aptitude, without which it might be difficult for her to reach a NS level of implicit knowledge. The fact that she was able to complete a doctoral degree in English is somewhat a reflection of her high level of language aptitude and L2 proficiency. In addition, the fact that she mainly used English in her social life indicates that she probably spent much more time using English for communication than those learners who only use English for work. In fact, as the NNS participants reported, the use of English in their personal lives was limited for the majority of them and they mainly relied on Mandarin (n = 72, see Table 4.66). This difference in daily language use could be a factor contributing to the different levels of implicit knowledge developed. Thus, it can be seen that the exceedingly successful case of participant 4 is very likely a result of the contribution of a multitude of different factors, including language aptitude and daily language

use.

The finding that language aptitude is a significant predictor of both implicit and explicit knowledge is contrary to Granena's (2014) conclusion that aptitude is not related to linguistic competence understood as implicit knowledge. In her study, language aptitude was found to be correlated with the scores in an untimed aural GJT but not in a timed aural GJT, and this led her to conclude that aptitude is only related to the explicit learning system. The present study and Granena (2014) both employed the LLAMA test to measure language aptitude, and the participants in both studies were long-term residents in the L2 country. However, the present study differs from Granena (2014) in terms of the number of tests that were used to measure L2 knowledge. The present study used four proposed tests of implicit knowledge, namely the EI, the aural TGJT, the WordM and the cloze test. More importantly, a rigorous process of test validation was conducted, and it was concluded that the EI and the TGJT were valid measures (See Sections 4.3 and 5.1), so that only the scores of these two tests were combined to form a score of implicit L2 knowledge (see Section 4.6). To measure explicit knowledge, Granena (2014) employed an aural untimed GJT, whereas the present study included three tests, including the written UGJT, the MKT and the cloze test based on the results of test validation (See Sections 4.3 and 5.1). These differences in how L2 knowledge is measured could be responsible for the different results between the present study and Granena (2014) regarding whether language aptitude is related to implicit knowledge. Specifically, since there is a lack of test validation in Granena (2014), the extent to which implicit and explicit knowledge is validly measured is in doubt. The use of an aurally presented untimed GJT could be actually tapping some implicit knowledge, because it is possible to use implicit knowledge to complete an untimed test, especially one that is presented aurally (see Section 5.1.2). Clearly, employing only one measure for each type of knowledge is a limitation.

The fact that age appeared as a significant predictor of explicit knowledge when language aptitude was controlled for was somewhat unexpected because previous

analysis has shown that age does not influence explicit knowledge (see Sections 4.9.2 and 5.4). However, since the significant age effect was only found in the late group (AoA > 30), it could be assumed that this was likely a result of the type of learning that this group of participants adopted. It is possible that these learners mainly relied on their explicit learning system even after immersion in the L2 country because the capacity of their implicit learning system was very limited. As a consequence, they had gradually developed a high level of explicit knowledge. In comparison, the younger learners who arrived in the L2 country before 30 years of age showed no significant differences in explicit knowledge, which could be seen as some evidence of the rather weak correlation between learning age and explicit knowledge. Therefore, it was concluded that language aptitude had a more important role in predicting the explicit knowledge of the NNS than age.

The failure to discover a significant interaction effect between age and language aptitude runs counter to the hypothesis that the influence of age on long-term L2 attainment is mediated by language aptitude. Rather, it was found that age and language aptitude are equally influential on implicit L2 knowledge, but language aptitude has a stronger influence on explicit knowledge than age. These findings emphasise the central importance of language aptitude in the development of both implicit and explicit knowledge irrespective of AoA. In turn, this finding implies that with an early immersion in the L2 context (i.e. early AoA) and a high level of language aptitude, there would be higher probabilities of success in the acquisition of the L2. Early immersion here could be as early as 3 years old, as was the case of participant 1 in the present study (see Table 4.61). The upper limit of early immersion could be up until puberty, as learners who were immersed after 13 years old in the present study all failed to be near-native (the exception being participant 4 as discussed earlier). In other words, late learners who have a high level of language aptitude may not be able to develop a high level of implicit L2 knowledge; and young learners whose level of language aptitude are low may not be able to do so, either. In other words, an early immersion in the L2 context and a high level of

language aptitude are both necessary conditions for a high level of implicit L2 knowledge.

The mechanism of how language aptitude facilitates the development of L2 knowledge is worth pondering. When age of immersion was controlled for, the significant role of language aptitude on implicit knowledge (see Section 4.10.1) provides some evidence that language aptitude is a pertinent factor for implicit learning. In other words, language aptitude affects the implicit knowledge of young learners by making an impact on their implicit learning system (see Section 5.5). This conclusion contradicts that of Granena (2014), who suggests that language aptitude has an influence on early learners, but its influence is only on the explicit learning system and not the implicit learning system. Granena's (2014) conclusion seems to be a paradox because early learners would probably rely on implicit learning since they are immersed in the L2 context and their implicit learning system is still active. Young learners would be unlikely to engage in explicit learning unless they are enrolled in intensive language courses. It is more likely for young learners to be enrolled in the general education system, such as primary and secondary schools, where English is used as the medium of instruction. Thus, it could be argued that if language aptitude plays a role for early learners, it has to be through the implicit learning system that its influenced is imposed. As previously discussed, the lack of sufficient and valid measures of implicit and explicit knowledge could be the reason for the differences in the conclusions between the present study and Granena (2014).

However, a word of caution is due here because implicit/explicit knowledge and implicit/explicit learning is not isomorphic (R. Ellis, 2015). It may seem intuitive to assume that implicit learning would result in implicit knowledge, but the actual learning and knowledge development processes are much more complicated. The interface issue is also relevant to the development of implicit knowledge, that is, the extent to which implicit knowledge converts into explicit knowledge and vice versa. It is possible that some explicit knowledge could be transferred to implicit

knowledge, as in the strong interface position suggested by some scholars (DeKeyser, 1998; 2007; Sharwood Smith, 1981). The present study was unable to unravel how exactly the implicit knowledge of the NNS was developed, especially regarding whether some explicit knowledge had been transferred to implicit knowledge. Nonetheless, it could be argued that implicit and explicit learning would primarily result in implicit and explicit knowledge, respectively. Thus, by examining the relationships amongst age, language aptitude, implicit and explicit knowledge, some inferences could be made in relation to implicit and explicit learning.

Nonetheless, given that language learning outcomes are subject to the influences of various other individual difference variables such as motivation, learning style and learning strategy, the amount of variance in implicit knowledge accounted for by age and language aptitude (42%, see Section 4.10.1) as two variables shows that these two variables are indeed two very influential factors. Specifically, language aptitude plays an equally important role as age on the development of implicit L2 knowledge. Comparatively speaking, language aptitude plays a more important role than age on the development of explicit L2 knowledge. It is therefore concluded that both age and aptitude are important factors that influence L2 learning outcomes. A higher level of aptitude would be advantageous for both early and adult learners.

In summary, the current results provide some evidence for the validity of the measures used in the present study. Research questions have been addressed using the present results in this chapter. In the next chapter, the main research findings will be summarised.

CHAPTER SIX

CONCLUSION

In this chapter, a summary of the main research findings will be presented first. Then, the theoretical implications are addressed, followed by a discussion of the methodological issues raised by the present study. Finally, the chapter concludes with some consideration of the limitations of the study and makes suggestions for further research directions.

6.1 Summary of main findings

RQ1: *What is the profiling of the implicit and explicit knowledge of the NNS and the NS in the present study? Is there a difference between the implicit and explicit knowledge of the NNS in comparison to the NS?*

Corresponding to this research question, it was predicted that the NNS would have more explicit knowledge than implicit knowledge, whereas the NS would have more implicit knowledge than explicit knowledge. Results show that this was indeed the case. This finding demonstrates that even after long-term residence in the L2 country, the level of explicit knowledge of the NNS is still higher than implicit knowledge. When compared to the NS, the NNS had a significantly lower level of implicit knowledge whereas they possessed a higher level of explicit knowledge.

These findings provide some evidence that implicit knowledge seems to be difficult to develop, even if learners have the opportunity of being immersed in a naturalistic context where there is abundant linguistic input of high quality. Given that the participants in the present study had all been immersed in the L2 country for more than 8 years and they had spent an average of 15.32 years in this country at the time of this study, the above findings seems to be in line with the claim that what can

be acquired implicitly by post-adolescent L2 learners in communicative contexts is quite limited (N. C. Ellis, 2008).

The overall lack of implicit knowledge of the group of NNS may perhaps be attributed to the fact that these participants continued to use their first language on a daily basis after immersion. As explained in Chapter 1 (see Section 1.2), there is a rather large Chinese population living in Auckland and Mandarin Chinese can be used by Chinese speakers for a wide range of social purposes. Based on the data collected from the questionnaire, it seems that many NNS participants indeed had more social interactions with other Chinese people and English was only used in their professional context. As a result of this, the NNS in the present study may have developed a restricted linguistic repertoire that was specific to their profession. However, to complete the language tests in the present study, the participants needed a wide linguistic repertoire, encompassing knowledge of a range of linguistic structures of different levels of complexity. It seems that the lack of English use outside the workplace of the NNS participants may have hindered the development of the kind of linguistic knowledge that the present study accessed. However, as self-reported data can be rather unreliable, this conclusion needs to be interpreted with caution.

RQ2: *What is the relationship between age, including age of onset and age of immersion in the L2 context, length of residence and implicit and explicit L2 knowledge?*

It was assumed that different age variables would have different influences on the implicit and explicit knowledge of the NNS. In the present study, two age variables were investigated. AO (i.e. age of onset) refers to the age that the learning of English begins, whereas AoA (i.e. age of immersion) refers to the age at which immersion in the L2 environment begins. In the case of the participants of the present study, the majority of them had different AO and AoA, AO indicated the start of learning English in China and AoA showed the age of immigration to New

Zealand. Results confirm that while AO (M = 10.15, SD = 3.05) does not have an influence on either implicit or explicit L2 knowledge, AoA (M = 21.40, SD = 10.10) has an influence on implicit knowledge but not on explicit knowledge. Length of residence in the L2 country was found to be unrelated to the levels of implicit and explicit knowledge of the NNS.

These findings provide some evidence that second language acquisition is maturationally constrained. However, if long-term learning outcomes are understood as implicit L2 knowledge, the present study shows that age of immersion is much more influential than age of onset. The pattern of influence of age of immersion is shown to be negatively linear, which runs counter to the postulation of a specific critical or sensitive period for second language acquisition.

The finding of the lack of correlation between length of residence and implicit and explicit knowledge indicates that differences in the amount and quality of input in different learning contexts do not suffice to account for the age effects in second language acquisition. It is likely that age influences acquisition outcomes by making an impact on the implicit learning system of learners, whereas the explicit learning system of learners remains unaffected. Parallel to the conclusions of DeKeyser (2003), this conclusion is in line with the original postulation made by Lenneberg (1967), which predicted that age would restrict learners' ability to learn in naturalistic contexts through mere exposure to the L2.

RQ 3: *What is the relationship between language aptitude, working memory and implicit and explicit L2 knowledge?*

The present study took a holistic view of language aptitude and working memory and calculated composite scores based on the individual scores of multiple sub-tests. Results show that language aptitude is influential for the development of both implicit and explicit knowledge, and higher levels of aptitude would lead to higher levels of implicit and explicit knowledge being developed. In comparison,

working memory was found to be not related to either implicit or explicit knowledge of the NNS.

The present study points out that making a distinction between implicit and explicit knowledge would provide a means to evaluate the role of language aptitude in relation to different learning contexts. The findings confirm that language aptitude indeed plays a very influential role in second language acquisition (de Graaff, 1997; Robinson, 1995a). The correlation between language aptitude and L2 outcomes indicates that language aptitude could be predictive of both learning rate and ultimate attainment. The stronger link between language aptitude and implicit knowledge in the present study further indicates that language aptitude could be of relevance to the type of learning that occurs in naturalistic learning contexts.

The lack of predictive power of working memory on L2 acquisition outcomes is an unexpected finding. However, it seems that this lack of correlation is a result of the cross-sectional nature of the present study, because it is ultimate attainment, rather than learning rate, that was measured. Thus, in line with previous research findings that suggest the relevance of working memory to learners with a low level of proficiency (Cheung, 1996; French, 2006) or to those at early L2 acquisition stages (Gass & Lee, 2011; Juffs & Harrington, 2011; Wen, 2016), the present study seems to have provided some evidence that the role of working memory diminishes as learners' proficiency level increases or as L2 acquisition proceeds to later stages.

However, it should be noted that the above findings are based on a limited number of a single population, that is, 86 Chinese learners of English. The extent to which the current findings could be generalised to other second language learners and learning contexts would require further investigation. It is possible that language aptitude plays an important role for Chinese learners of English because these two languages are distinctively different as they come from different language families. Chinese is a Sino-Tibetan language whereas English is an Indo-European language, which is probably why many Chinese learners generally find it very difficult to learn English. However, a similar degree of difficulty may not be experienced by learners

whose first language belongs to the same language family as English, for example, German.

The other aspect of the present findings that calls for cautious interpretation is concerned with the validity of the language aptitude test. The present study used the LLAMA test (Meara, 2005), a test that remains to be validated further. It seems that researchers tend to disagree with regard to the validity of language aptitude measures, and this is partly because there is considerable controversy over the components that the construct of language aptitude entails. For example, the relevance of working memory to language aptitude remains an area of ongoing investigation (see Section 6.3.2 for further discussion).

RQ4: *What is the joint influence of age and language aptitude on implicit and explicit L2 knowledge? Is there an interaction between age and aptitude?*

The joint influence of age and language aptitude and their interaction have been examined in relation to implicit and explicit L2 knowledge. Results show that while age and aptitude are robust predictors of L2 outcomes independently, there is a lack of evidence for their interaction. It was found that both age and aptitude are significant predictors of implicit L2 knowledge, while language aptitude is the only significant predictor of explicit L2 knowledge.

The equally influential impact of both age and language aptitude on implicit knowledge indicates that for the NNS in the present study, an early age of immersion and a high level of aptitude are the necessary conditions for the development of a high level of implicit knowledge. In other words, an early age of immersion would be an advantage for learners with the same level of language aptitude, and a higher level of language aptitude would be advantageous for learners of the same age of immersion. Analysis of the near-native NNS learners provides further evidence in this respect, as the results show that the majority of those who were able to score within the range of NS in implicit knowledge also had an above average level of

language aptitude. In addition, the majority of those who were immersed in the L2 environment before puberty (i.e. ≤ 13 years old) but failed to score within the NS range in implicit knowledge only had an average level of language aptitude. Based on these findings, it seems that, for the Chinese learners of English in the present study, immersion in the target language environment before the age of puberty and an above average level of language aptitude are both necessary for near-native L2 acquisition outcomes.

The failure to find a significant interaction effect of age and language aptitude shows that language aptitude does not work as a mediating factor of age, which contradicts the prediction of the fundamental difference hypothesis (Bley-Vroman, 1988; 1989; 2009) and some previous research findings (DeKeyser, 2000; DeKeyser et al., 2010). As stated in the previous paragraph, the findings of the present study show that a higher level of language aptitude confers an advantage to both the early and the adult L2 learners of this study.

In summary, this study set out to investigate the extent to which age and language aptitude influence the development of implicit and explicit L2 knowledge. It is concluded that:

1) Age of immersion, that is, the age at which exposure to the target language in a naturalistic context begins, influences the long-term implicit language knowledge but not the explicit knowledge of the NNS. This conclusion is demonstrated by the significant predictive power of age of immersion on implicit knowledge rather than on explicit knowledge. In comparison, age of onset, that is, the age at which learning begins in an instructed context, does not influence either implicit or explicit knowledge of the NNS. It is further concluded that the influence of age on second language acquisition is likely realised by making an impact on the implicit learning system of learners.
2) Language aptitude, defined as a set of cognitive abilities in language acquisition, has an influence on both the implicit and explicit knowledge of

the NNS. This influence is evidenced by the significant predictive power of language aptitude on both implicit and explicit knowledge in the regression analyses. This finding indicates that language aptitude could be pertinent to both implicit and explicit learning systems. Working memory, on the other hand, does not seem to influence long-term implicit and explicit knowledge.

3) For the Chinese learners of English in this study, a young age of immersion and a high level of language aptitude are both necessary conditions for developing a high level of implicit knowledge. This conclusion is evidenced by the significant predictive power of both age and language aptitude in the regression analysis that examined the joint influence of these two factors. In turn, this finding indicates that a high level of language aptitude is not only beneficial for adult learners; it is also an advantageous factor that could assist early/child learners of L2.

6.2 Theoretical implications

In terms of the influence of learning age, the negative linear correlation found between age of immersion and implicit knowledge supports a general maturational constraints view of second language acquisition, and does not favour the hypothesis of (a) particular critical/sensitive period(s). Although the findings of the present study seem to indicate that pre-puberty age would be advantageous for the Chinese learners of English under investigation, a gradual decrease in the level of implicit knowledge was observed as learners' age of immersion increased. This negative linear pattern was evident even when learners were separated into different age groups (see Table 4.53 and Figure 4.12), and this finding further supports the notion of maturational constraints in second language acquisition. If the form of age effects was manifested as a critical/sensitive period, there would have been a point of change, after which the level of implicit knowledge levelled off. Thus, the failure to identify such a point of change, suggests that second language acquisition outcomes

are constrained by the maturational factor, namely age of immersion in the L2 context, although the notion of (a) critical/sensitive period(s) may not be maintained.

Nonetheless, the present research findings lend strong support to the robustness of the age variable in second language acquisition. By testing implicit and explicit knowledge separately, the present results indicate that there could be a big gap between the implicit knowledge of second language learners and native speakers as far as implicit knowledge is concerned. In turn, this finding appears to corroborate the prediction made by Lenneberg (1967) that learning age affects the implicit learning system only. As learner's age increases, it is likely that the implicit learning system gradually loses power, and therefore leads to a decrement in the implicit knowledge that results from implicit learning. This putative association between learning age and implicit learning corroborates the findings in psychological studies, which also show that implicit learning is sensitive to the influence of age (Fletcher, Maybery, & Bennett, 2000; Maybery, Taylor, & O'Brien-Malone, 1995).

The finding that language aptitude is predictive of long-term L2 outcomes lends support to the robustness of language aptitude in second language acquisition in general. In addition, the finding that language aptitude has an ancillary role for the entire group of participants in the present study appears to be indicative of the relevance of language aptitude to both implicit and explicit learning. As is shown by the current results, it is clear that, like age of immersion, language aptitude is identified as a significant predictor of the implicit knowledge of the NNS (see Table 4.54). This means that unlike previous claims that the construct of language aptitude is largely explicit in nature (Li, 2015), there might be aspects of language aptitude that are also sensitive to implicit learning processes (Robinson, 1995a; 2001). This conclusion is therefore in keeping with the postulation made by Skehan (1986; 1989; 1998; 2012) that, language aptitude should play an equally important role in implicit learning as it does in explicit learning because it reflects a general ability of input processing. Nonetheless, as noted earlier, the relevance of language aptitude to implicit learning and implicit knowledge is in need of further investigation as the

validity of language aptitude measures remains an empirical issue.

Considering the important roles of language aptitude and working memory in second language acquisition, several researchers have argued for the incorporation of working memory as a language aptitude component (Skehan, 2015b; Wen & Skehan, 2011). The extent to which working memory should be included as an aptitude component is explored in the present study by investigating the influence of language aptitude and working memory ability on L2 outcomes. The results show that in comparison with language aptitude, the association between working memory ability and long-term L2 outcomes of the NNS is minimal. However, this lack of association by no means undermines the role of working memory in second language acquisition. Rather, this finding seems to provide some evidence for the hypothesis that working memory is involved at different second language acquisition stages (Skehan, 2002, 2012, 2015b). Following Skehan's (2002, 2012) hypothesis, working memory plays a more important role at initial stages of acquisition, including input processing, noticing and handling form and meaning simultaneously, and pattern identification, in comparison with later acquisition stages such as automatization and lexicalization. As the present study measured the long-term L2 outcomes that were clearly beyond the initial acquisition stages, the rather weak correlation between working memory and L2 outcomes appears to be in line with the theoretical predictions made by Skehan (2002, 2012). Similarly, the finding that the language aptitude test used, that is, the LLAMA test was measuring a different construct with the two working memory tests, namely the non-word repetition test and the listening sentence span test should not be taken as evidence to oppose the idea of considering working memory as an aptitude component. Rather, the fact that the factor analysis results clearly demonstrate a two-factor model (see Section 5.2.3) highlights the need for further research, so as to understand how working memory and language aptitude are involved in second language acquisition.

As the primary aim of the present study is to explain the influences of age and language aptitude on long-term second language acquisition outcomes, direct

pedagogical implications are not provided. However, at a macro level, the lack of predictive power of age of learning in instructed contexts seems to indicate that an early start in such contexts would not lead to better learning achievements. In order to promote second language acquisition in instructed contexts, it is probably essential for the teachers to create a learning environment that resembles naturalistic learning contexts, where linguistic input is of high quality and abundant. Equally important is the need for engaging learners in communicative tasks in the L2 classrooms, so as to help the learners develop some competence in applying their learnt knowledge in language production.

Interestingly, it seems that functioning effectively in an L2 environment does not necessarily imply that a learner has a high level of implicit L2 knowledge. As reported by the NNS participants in the present study, they were able to use their L2 for work and living in New Zealand. However, at a linguistic level, there was a lack of implicit knowledge for the majority of these participants because they were unable to intuitively identify the grammatical errors in the language tests or to produce error-free sentences under time pressure. Considering the fact that the NNS in the present study had received a considerable amount of English instruction that can probably be characterised as *focus on forms* (i.e. deliberate teaching of grammar), the lack of accuracy in the language tests of these participants seems to show that such instruction makes little contribution to learners' correctness in communication (i.e. their implicit knowledge). Thus, pedagogically speaking, perhaps a better instruction method in language classrooms would be to *focus on form*, by bringing grammar to the attention of learners as part of communicative language practice (Doughty, 2003; R. Ellis, Basturkmen, & Loewen, 2001; Erlam & Ellis, 2018a; 2018b; Long, 1991), so that learners would be offered opportunities to notice the gap between their inter-language and the target language. Eventually, this may help learners with the development of higher levels of implicit knowledge.

6.3 Methodological issues raised by research

6.3.1 Difficulty of measuring implicit and explicit knowledge

The present study once again highlights the difficulty of designing reliable measures of implicit and explicit knowledge. As discussed in Chapter 5 (see Section 5.1), accessing implicit knowledge could be challenging as meticulous preparation needs to be done at various stages, including test material design, test delivery modality, test administration, and scoring. Arguably, tests that require spontaneous use of L2 knowledge are more likely to require a primary reliance on implicit knowledge, rather than explicit knowledge. This could be achieved empirically by directing the attention of the participants away from language form, including both grammatical and ungrammatical stimuli sentences, and imposing a time pressure in tests (R. Ellis, et al., 2009). However, it is particularly challenging to decide how much time pressure is appropriate so that the participants would not have the opportunity to resort to their explicit knowledge but would be able to provide an answer. For example, in the present study, a 10 second repetition time was imposed for sentence repetition in addition to the original requirement of listen once and in real time by Erlam (2006) in the elicited imitation test. Erlam (2006) did not impose a time pressure for sentence repetition, but the present study did so with the aim of increasing the likelihood of drawing on implicit knowledge. There is some evidence in the present study that this further time pressure seems to have led to the use of implicit knowledge (see Section 5.1.1), but the extent to which longer or shorter time pressure would be similarly influential is not addressed. It seems that researchers are left to their own decisions with regard to the specific requirement of time pressure in research practice, and therefore some degree of arbitrariness is always involved in this process. By the same token, test administration in elicited imitation tests is of equal importance as the manipulation of time pressure, in the sense that researchers need to be very careful with test instructions so that the attention of participants are not derived to linguistic form (see Section 5.1.1). Similarly, aspects such as test

modality and time pressure in timed grammaticality judgement tests could be investigated further. For timed grammaticality judgement tests in general, researchers once again are confronted with the decision of making appropriate time restrictions for test completion. Further research is also needed to investigate the influence of test modality, as stimuli that are presented aurally or visually may tap into different types of knowledge in the timed grammaticality judgement tests. Nonetheless, it could be argued that the possibility of accessing implicit knowledge can be maximised depending on the synergy of multiple efforts made to test design and test administration, although it is not possible to design pure measures of implicit knowledge (R. Ellis, 2005).

Measuring explicit L2 knowledge is as challenging as measuring implicit knowledge. Due to the systematic and automatic nature of implicit knowledge, it seems impossible to design tests to access explicit knowledge that are entirely devoid of implicit knowledge if both types of knowledge are available. In other words, as tests that are designed as measures of explicit knowledge could be completed with implicit knowledge, it is difficult for the researcher to make accurate decisions regarding which type of knowledge is involved. Asking the participants to report the source of judgement (i.e. by "rule" or by "feel") in untimed grammaticality judgement tests, could provide some hints to some extent (Schmidt, 1994); but it should also be noted that self-reported data often has low reliability.

On the other hand, tests that measure metalinguistic knowledge, such as the metalinguistic knowledge test used in the present study, are undoubtedly accessing explicit knowledge. However, albeit being explicit in nature, it is generally agreed that one's metalinguistic knowledge is independent of his or her explicit knowledge (R. Ellis, 2004). This means that the metalinguistic knowledge tests can only access explicit knowledge in part. To best assess one's explicit knowledge, then, would require multiple measures including a metalinguistic knowledge test. Yet, as argued earlier it is difficult to exclude the possibility of the involvement of implicit knowledge in other tests of explicit knowledge, such as untimed grammaticality

judgement tests. This seems to be an empirical dilemma and suggests the need for more research in developing valid measures of explicit knowledge.

6.3.2 Difficulty of measuring language aptitude and working memory

As discussed in Chapter 5 (see Section 5.2), there is a lack of validation studies of the current most widely used measure of language aptitude, that is, the LLAMA test. The LLAMA test is loosely based on the MLAT (Carroll & Sapon, 1959), which seems to have the tendency of involving explicit aptitude components, such as language analytic ability, for test completion. It has been suggested that LLAMA D could be measuring an implicit aptitude component (Granena, 2012; 2013b), but whether this claim is true is unclear so far considering the existing counter evidence provided by Rogers et al. (2017) and the present study, both of which showing that this sub-test measures the same construct as the other three sub-tests. Clearly, not until the LLAMA test is validated further can researchers draw conclusions about what aptitude components are the possible contributors to implicit and explicit L2 knowledge. Due to this, the present finding that language aptitude is associated with both implicit and explicit L2 knowledge needs to be interpreted with caution.

For the tests of working memory, there is considerable difficulty in test scoring. Although the non-word repetition test and the listening sentence span test have been widely used, researchers vary in how these tests are scored. For example, the present study used the percentage method by dividing the number of correctly repeated non-words by the total number of nonwords in each trial and then averaging these out across all trials (see Section 4.2). Alternatively, one could count the total number of correctly repeated nonwords or count the number of words that are correctly repeated in each trial (maximum set score). These multiple options are left to the researcher's decision, so that there is a chance that personal bias is introduced into the research. This issue also occurs in the scoring of the sentence span test.

6.4 Limitations

The present study has a number of weaknesses. Firstly, as a cross-sectional study that aims to explain the influences of two individual difference factors (i.e. language aptitude and age) on long-term second language acquisition outcomes, it could benefit from a larger sample size. Involving more NNS participants would be particularly beneficial for conducting confirmatory factor analysis and would allow for a more precise examination of the relationships among theoretical constructs. Secondly, although the involvement of a single learner population (i.e. Chinese learners of English) in the present study has the advantage of controlling for the possible influence of L1, the generalisability of the present findings to other learner populations is restricted. The robust role of language aptitude as identified in the present study could be due to the fact that the L1 and L2 of this group of learners are distinctively different, so that learners who have higher cognitive ability show an advantage over the others in the acquisition of a very different language. Thirdly, the limited number of tests of working memory might have prevented the discovery of a relationship between working memory and L2 outcomes. Had other measures of working memory been employed, this construct could have been measured more rigorously and so the working memory – L2 outcomes relationship could have been examined. Lastly, the failure to find early NNS participants (AoA ≤ 13) and late NNS participants (AoA > 30) who have an excellent level of language aptitude might be responsible for the lack of evidence for the possible interaction effect of age and language aptitude. However, given the sampling strategy of the present research, such learners are very difficult to find.

6.5 Future research directions

There is a need to further investigate the mechanisms of how age and language aptitude influence long-term second language acquisition outcomes, in terms of the

distinction between implicit and explicit knowledge. Involving a larger number of participants in future research would be beneficial. A large sample size would not only better reflect individual variations in long-term second language learning outcomes, but also lay the foundation for rigorous statistical analysis using advanced data analysis methods such as factor analysis. Moreover, future research would also benefit from involving learners of a different L1- L2 pairing, for example German learners of English or English learners of Mandarin Chinese. Investigations using a greater range of learner populations would further examine the robustness of age and language aptitude, therefore enabling researchers to evaluate the relevance of these two factors to second language acquisition in general.

The relevance of language aptitude to implicit learning and implicit knowledge is an area that could be explored further. To examine the relationship between language aptitude, implicit knowledge and implicit learning, future research could specifically target early learners who mainly acquire their L2 in naturalistic contexts (e.g. age of immersion of 13 years old or earlier) with minimal or no experience of explicit learning. Investigations of such learners could probably provide some evidence for the relevance of language aptitude in implicit learning conditions. However, as previously noted, the validity of measures, including measures of L2 outcomes and language aptitude, is also of central importance.

There is a need for further research with respect to the validation of language aptitude tests. The current most popular aptitude test, the LLAMA test needs to be validated further. Validation of the LLAMA test would not only bring transparency over what language aptitude components are accessed by this test, but also make progress in the theoretical understanding of the construct of language aptitude in general. Of course, if researchers are more confident about the kind of cognitive abilities that language aptitude tests measure, more reliable conclusions could be made regarding the role of these cognitive abilities in language acquisition.

There is also a need for additional research that investigates the relationship between language aptitude and working memory. Specifically, future research could

take the direction of investigating the relevance of working memory to different L2 acquisition stages, so as to get a better understanding of how working memory works in second language acquisition. Studies that include both language aptitude measures and working memory measures could also potentially examine the relationship between the two, and therefore provide evidence for the proposal of including working memory as an aptitude component.

With regard to the measurement of implicit and explicit knowledge, further research is needed to ascertain whether tests that measure reaction times, such as the word monitoring test, are indeed valid measures of implicit knowledge. There is also a need to understand how these measures relate to the existing measures of implicit knowledge, including the oral elicited imitation tests and the timed grammaticality judgement tests. Similarly, further research is needed to improve current measures of implicit and explicit knowledge. For example, developing measures of level of consciousness that are more accurate than asking participants to report their source of judgement in the untimed grammaticality judgement tests would probably allow for more accurate interpretation of the type of knowledge being accessed by this test.

REFERENCES

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition, 34*(2), 187-214.

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*(4), 481-509.

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249-306.

Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta Psychologica, 128*(3), 466-478.

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics, 20*(3), 242-275.

Agresti, A. (2002). *Categorical data analysis*. New York: John Wiley & Sons.

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research, 1*(2), 93-121.

Alptekin, C., & Erçetin, G. (2011). Effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly, 45*(2), 235-266.

Andringa, S. (2014). The use of native speaker norms in critical period hypothesis research. *Studies in Second Language Acquisition, 36*(3), 565-596.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556-559.

Baddeley, A. (1998). Recent developments in working memory. *Current Opinion in Neurobiology, 8*(2), 234-238.

Baddeley, A. (1999). *Essentials of human memory.* Hove England: Psychology Press.

Baddeley, A. (2007). *Working memory, thought, and action.* Oxford: Oxford University Press.

Baddeley, A., & Gathercole, S. (1992). Learning to read: The role of the phonological loop. In J. Alegria, D. Holender, J. Junca de Morais & M. Radeau (Eds.), *Analytic approaches to human cognition* (pp. 153-167). Amsterdam: North-Holland.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158-173.

Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47-89.

Bates, D., Machler, M., & Bolker, B. (2017). *Lme4: Linear mixed-effects models using 'Eigen' and s4.* R package version 1.1-15. URL https://cran.r-project.org/web/packages/lme4/index.html.

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*(1), 78-117.

Bialystok, E., & Hakuta, K. (1994). *In other words: The psychology and science of second language acquisition*. New York: Basic Books.

Bialystok, E., & Hakuta, K. (1999). Confounded age: Linguistic and cognitive factors in age differences for second language acquisition. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 161-181). Mahwah, NJ: Lawrence Erlbaum.

Bialystok, E., & Miller, B. (1999). The problem of age in second-language acquisition: Influences from language, structure, and task. *Bilingualism: Language and Cognition, 2*(2), 127-145.

Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language, 68*, 706-755.

Birdsong, D. (1999). *Second language acquisition and the critical period hypothesis*. Mahwah, NJ: Lawrence Erlbaum.

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning, 56*(s1), 9-49.

Birdsong, D. (2009). Age and the end state of second language acquisition. In T. K. Bhatia (Ed.), *The new handbook of second language acquisition* (pp. 401-424). Bingley: Emerald Group Publishing Ltd.

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language, 44*, 235-249.

Bley-Vroman, R. (1988). The fundamental character of foreign language learning. In W. Rutherford, & M. Sharwood Smith (Eds.), *Grammar and second language teaching: A book of readings* (pp. 19-30). Rowley, MA: Newbury House.

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning. *Linguistic Perspectives on Second Language Acquisition, 4*, 1-68.

Bley-Vroman, R. (2009). The evolving context of the fundamental difference hypothesis. *Studies in Second Language Acquisition, 31*(2), 175-198.

Bolibaugh, C., & Foster, P. (2013). Memory-based aptitude for nativelike selection: The role of phonological short-term memory. In G. Granena, & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 203-228). Amsterdam: John Benjamins Publishing.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 133-159). Mahwah, NJ: Lawrence Erlbaum.

Bongaerts, T., Mennen, S., & van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of dutch as a second language. *Studia Linguistica, 54*(2), 298-308.

Bongaerts, T., Planken, B., & Schils, E. (1995). Can late starters attain a native accent in a foreign language? A test of the critical period hypothesis. In D. Singleton, & Z. Lengyel (Eds.), *The age factor in second language acquisition* (pp. 30-50). Clevedon: Multilingual Matters.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition, 19*(4), 447-465.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 33*(2), 247-271.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

Butler, Y. G. (2000). The age effect in second language acquisition: Is it too late to acquire native-level competence in a second language after the age of seven. In Y. Oshima-Takane, Y. Shirai & H. Shirai (Eds.), *Studies in language sciences* (pp. 159-169). Tokyo: The Japanese Society for Language Sciences.

Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2009). The role of language aptitude in first language attrition: The case of pre-pubescent attriters. *Applied Linguistics, 31*(3), 443-464.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.) Routledge.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83-118). Rowley, MA: Newbury House.

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test*. San Antonio, TX: Psychological Corporation.

Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary language. *Developmental Psychology, 32*(5), 867-873.

Child, J. R. (1998). Language aptitude testing: Learners and applications. *Applied Language Learning, 9*, 1-10.

Clapham, C. (2001). The assessment of metalinguistic knowledge. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, . . . K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 33-43). Cambridge: Cambridge University Press.

Coolidge, F. L., & Wynn, T. (2009). *The rise of homo sapiens: The evolution of modern thinking*. Oxford: Oxford University Press.

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language, 63*(3), 544-573.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching, 5*(1-4), 161-169.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Cowan, N. (2005). *Working memory capacity*. New York: Psychology Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Publishers.

Curtiss, S. (1977). *Genie: Psycholinguistic study of a modern-day "wild child"*. London: Academic Press.

Curtiss, S. (1988). Abnormal language acquisition and the modularity of language. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge survey: Volume 2, linguistic theory: Extensions and implications* (pp. 96-116). Cambridge: Cambridge University Press.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422-433.

Danks, J. (1980). Comprehension in listening and reading: Same or different? In J. Danks, & K. Pezdek (Eds.), *Reading and understanding* (pp. 1-39). Newark, DE: International Reading Association.

Davies, A. (1991). *The native speaker in applied linguistics*. Edinburgh: Edinburgh University Press.

de Graaff, R. (1997). The "eXperanto" experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition, 19*(2), 249-276.

DeKeyser, R. M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42-63). Cambridge: Cambridge University Press.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 499-533.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313-348). Oxford: Blackwell Publishing Ltd.

DeKeyser, R. M. (2007). Introduction: Situating the concept of practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1-18). Cambridge: Cambridge University Press.

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning, 63*(s1), 52-67.

DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics, 31*, 413-438.

DeKeyser, R. M., & Koeth, J. (2011). Cognitive aptitudes for L2 learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning: Volume II* (pp. 395-406). New York: Routledge.

DeKeyser, R. M., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll, & A. M. B. de Groot (Eds.), *Handbook of bilinguialism: Psycholinguistic approaches* (pp. 88-108). Oxford: Oxford University Press.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22*(5), 735-808.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1-11.

Dörnyei, Z. (2010). The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective. In E. Macaro (Ed.), *The Bloomsbury companion to second language acquisition* (pp. 247-267). London: Bloomsbury.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589-630). Oxford: Blackwell Publishing Ltd.

Doughty, C. J. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 256-310) Blackwell Publishing Ltd.

Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal, 79*(1), 67-89.

Eisenstein, M. (1980). Grammatical explanations in ESL: Teach the student, not the method. *TESL Talk, 2*, 3-13.

Elder, C. (2009). Validating a test of metalinguistic knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 113-166). Bristol: Multilingual Matters.

Elder, C., & Manwaring, D. (2004). The relationship between metalinguistic knowledge and learning outcomes among undergraduate students of Chinese. *Language Awareness, 13*(3), 145-162.

Elder, C., Warren, J., Hajek, J., Manwaring, D., & Davies, A. (1999). Metalinguistic knowledge: How important is it in studying a language at university? *Australian Review of Applied Linguistics, 22*(1), 81-96.

Ellis, N. C. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology Section A, 49*(1), 234-250.

Ellis, N. C. (1994). *Implicit and explicit learning of languages*. London: Academic Press.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*(2), 305-352.

Ellis, N. C. (2008). Implicit and explicit knowledge about language. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (pp. 1878-1890). New York: Springer.

Ellis, R. (1993). Second language acquisition and the structural syllabus. *TESOL Quarterly, 27*(1), 91-113.

Ellis, R. (1994). A theory of instructed second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 79-114). London: Academic Press.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning, 54*(2), 227-275.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*(2), 141-172.

Ellis, R. (2006). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics, 27*, 431-463.

Ellis, R. (2008a). Explicit form-focused instruction and second language acquisition. In B. Spolsky, & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 437-455). Malden, MA: Blackwell Publishing Ltd.

Ellis, R. (2008b). *The study of second language acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 3-25). Bristol: Multilingual Matters.

Ellis, R. (2015). *Understanding second language acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Preemptive focus on form in the ESL classroom. *TESOL Quarterly, 35*(3), 407-432.

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition, 29*(1), 119-126.

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol: Multilingual Matters.

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research, 9*(2), 147-171.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464-491.

Erlam, R. (2009). The elicited oral imitation test as a measure of implicit knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 65-93). Bristol: Multilingual Matters.

Erlam, R. & Akakura, M. (2016). New development in the use of elicited imitation. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 93-111). New York: Routledge.

Erlam, R. & Ellis, R. (2018a). Task-based language teaching for beginner-level learners of L2 French: An exploratory study. *The Canadian Modern Language Review, 74*(1), 1-26.

Erlam, R. & Ellis, R. (2018b). Input-based tasks for beginner-level learners: An approximate replication and extension of Erlam & Ellis (2018). *Language Teaching.* Advanced online publication. doi:10.1017/S0261444818000216.

Eubank, L., & Gregg, K. R. (1999). Critical periods and (second) language acquisition: Divide et impera. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 65-99). Mahwah, NJ: Lawrence Erlbaum.

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Felix, S. W. (1985). More evidence on competing cognitive systems. *Second Language Research, 1*, 47-72.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage publications.

Flege, J. E. (1999). Age of learning and second language speech. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 101-131). Mahwah, NJ: Lawrence Erlbaum.

Flege, J. E., & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition, 23*, 527-552.

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /l/ and /l/ accurately. *Language and Speech, 38*(1), 25-55.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second language acquisition. *Journal of Memory and Language, 41*, 78-104.

Fletcher, J., Maybery, M. T., & Bennett, S. (2000). Implicit learning differences: A question of developmental level? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 246-252.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*(1), 116-124.

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition, 36*(1), 101-132.

Fox, J., Friendly, M., & Monette, G. (2017). *Heplots: Visualizing tests in multivariate linear models.* R package version 1.3-4. URL https://CRAN.R-project.org/package=heplots

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). London: Sage Publications.

French, L. M. (2006). *Phonological working memory and second language acquisition: A developmental study of francophone children learning English in Quebec*. New York: Edwin Mellen Press.

Friendly, M. (2007). HE plots for multivariate linear models. *Journal of Computational and Graphical Statistics, 16*(2), 421-444.

Friendly, M. (2010). HE plots for repeated measures designs. *Journal of Statistical Software, 37*(4), 1-37.

Ganschow, L., & Sparks, R. (1995). Effects of direct instruction in Spanish phonology on the native-language skills and foreign-language aptitude of at-risk foreign-language learners. *Journal of Learning Disabilities, 28*(2), 107-120.

Gass, S., & Lee, J. (2011). Working memory capacity, inhibitory control, and proficiency in a second language. In M. S. Schmid, & W. Lowie (Eds.), *Modeling bilingualism: From structure to chaos.in honor of Kees de Bot* (pp. 59-84). Amsterdam: John Benjamins Publishing.

Godfroid, A., Loewen, S., Jung, S., Park, J., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge. *Studies in Second Language Acquisition, 37*(2), 269-297.

Granena, G. (2012). *Age differences and cognitive aptitudes for implicit and explicit learning in ultimate second language attainment* (Unpublished doctoral dissertation). University of Maryland, College Park.

Granena, G. (2013a). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In G. Granena, & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105-130). Amsterdam: John Benjamins Publishing.

Granena, G. (2013b). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63*(4), 665-703.

Granena, G. (2014). Language aptitude and long-term achievement in early childhood L2 learners. *Applied Linguistics, 35*(4), 483-503.

Granena, G. (2016). Elicited imitation as a measure of implicit L2 knowledge. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 185-204). Amsterdam: John Benjamins Publishing.

Granena, G., & Long, M. H. (2013a). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311-343.

Granena, G., & Long, M. H. (2013b). *Sensitive periods, language aptitude, and ultimate L2 attainment*. Amsterdam: John Benjamins Publishing.

Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioral sciences* (10th ed.). Boston, MA: Cengage Learning.

Green, D. (2003). Neural basis of the lexicon and the grammar in L2 acquisition: The convergence hypothesis. In R. van Hout, A. Hulk, F. Kuiken & R. J. Towell (Eds.), *The lexicon-syntax interface in second language acquisition* (pp. 197-218). Amsterdam: John Benjamins Publishing.

Green, P. S., & Hecht, K. (1992). Implicit and explicit grammar: An empirical study. *Applied Linguistics, 13*(2), 168-184.

Gregg, K. R. (2003). The state of emergentism in second language acquisition. *Second Language Research, 19*(2), 95-128.

Grigornko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal, 84*(3), 390-405.

Groth-Marnat. (2003). *Handbook of psychological assessment* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition, 35*(3), 423-449.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis: Pearson new international edition* (7th ed.). Harlow: Pearson Education.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science, 14*(1), 31-38.

Halliday, M. A., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research, 2*(1), 1-23.

Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition, 19*(3), 379-400.

Harley, B., & Hart, D. (2002). Age, aptitude, and second language learning on a bilingual exchange. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 301-330). Amsterdam: John Benjamins Publishing.

Harley, B., & Wang, W. (1997). The critical period hypothesis: Where are we now. In De Groot, A. M. B., & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 19-51). Mahwah, NJ: Lawrence Erlbaum.

Harrell, F. E. J. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival anaysis* (2nd ed.). New York: Springer.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14*(1), 25-38.

Hernandez, A., & Li, P. (2007). Age of acquisition: Its neural and computational mechanisms. *Psychological Bulletin, 133*(4), 638-650.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences, 9*(5), 220-225.

Hess, E. H. (1975). The ethological approach to socialization. In J. Sants, & H. J. Butcher (Eds.), *Developmental psychology: Selected readings* (pp. 57-74). Harmondsworth: Penguin Books.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. *Teaching collocation: Further development in the lexical approach* (pp. 47-69). Oxford: Oxford University Press.

Hu, G. (2005). English language education in china: Policies, progress, and problems. *Language Policy, 4*, 5-24.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Huang, B. H. (2016). A synthesis of empirical research on the linguistic outcomes of early foreign language instruction. *International Journal of Multilingualism, 13*(3), 257-273.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Hulstijn, J. H. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research, 18*(3), 193-223.

Hulstijn, J. H. (2010). Measuring second language proficiency. In E. Blom, & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 185-199). Amsterdam: John Benjamins Publishing.

Hwu, F., & Sun, S. (2012). The aptitude-treatment interaction effects on the learning of grammar rules. *System, 40*(4), 505-521.

Hyltenstam, K., & Abrahamsson, N. (2000). Who can become native-like in a second language? all, some, or none? *Studia Linguistica, 54*(2), 150-166.

Hyltenstam, K., & Abrahamsson, N. (2001). Age and L2 learning: The hazards of matching practical "implications" with theoretical "facts". (Comments on Stefka H. Marinova-todd, D. Bradford Marshall, and Catherine E. Snow's "Three misconceptions about age and L2 learning"). *TESOL Quarterly, 35*(1), 151-170.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 538-588). Oxford: Blackwell Publishing Ltd.

Hyltenstam, K. (1992). Non-native features of near-native speakers: On the ultimate attainment of childhood L2 learners. *Advances in Psychology, 83*, 351-368.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers*. Milwaukee, WI: ASQC Quality Press.

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis. *Studies in Second Language Acquisition, 16*(1), 73-98.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the *N:q* hypothesis. *Structural Equation Modeling, 10*, 128-141.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434-446.

James, C., Garrett, P., & Candlin, C. N. (2014). *Language awareness in the classroom*. London: Routledge.

Jarek, S. (2012). *Mvnormtest: Normality test for multivariate variables.* R package version 0.1-9. URL https://CRAN.R-project.org/package=mvnormtest.

Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 131-158). New York: Routledge.

Jia, G. (1998). *Beyond brain maturation: The critical period hypothesis in second language acquisition revisited* (Unpublished doctoral dissertation). New York University, New York.

Jia, G., & Fuse, A. (2007). Acquisition of English grammatical morphology by native mandarin-speaking children and adolescents: Age-related differences. *Journal of Speech, Language, and Hearing Research, 50*(5), 1280-1299.

Jiang, N. (2013). *Conducting reaction time research in second language studies*. New York: Routledge.

Johnson, J. S. (1992). Critical period effects in second language acquisition: The effect of written versus auditory materials on the assessment of grammatical competence. *Language Learning, 42*(2), 217-248.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*(1), 60-99.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching, 44*(2), 137-166.

Kachiniske, I., & Vafaee, P. (2014). Reexamining the validity of GJTs as measures of implicit knowledge: Reanalysis of Ellis & Loewen (2007), Bowles (2011), and Gutiérrez (2013). Paper presented at the Second Language Research Forum, University of South Carolina, Columbia, South Carolina.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486-507.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*(1), 165-172.

Kilborn, K., & Moss, H. (1996). Word monitoring. *Language and Cognitive Processes, 11*(6), 689-694.

Kim, J., & Nam, H. (2017). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition, 39*(3), 431-457.

Klein, W. (1995). The acquisition of English. In R. Dietrich, W. Klein & C. Noyau (Eds.), *The acquisition of temporality in a second language* (pp. 31-70). Amsterdam: John Benjamins Publishing.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.*).* New York: The Guilford Press.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal, 6*(2), 151-162.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition, 11*(2), 261-271.

Krashen, S. D., Long, M. H., & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition. *TESOL Quarterly, 13*, 573-582.

Krashen, S. D. (1981). *Second language acquisition and second language learning* (1st ed.). Oxford: Oxford University Press.

Krashen, S. D. (1982). *Principles and practice in second language learning*. Oxford: Pergamon Press Inc.

Larson-Hall, J. (2001). *Language acquisition by Japanese speakers: Explaining the why, how and when of adult learners' segmental success* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh.

Lawler, J., & Selinker, L. (1971). On paradoxes, rules, and research in second language learning. *Language Learning, 21*(1), 27-43.

Lecumberri, M. G., & Gallardo, F. (2003). English FL sounds in school learners of different ages. In M. Mayo, & M. G. Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 115-135). Clevedon: Multilingual Matters.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

Leung, J., & Williams, J. N. (2011). The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition, 33*(1), 33-55.

Lewandowsky, S., Oberauer, K., Yang, L., & Ecker, U. K. (2010). A working memory test battery for MATLAB. *Behavior Research Methods, 42*(2), 571-585.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal, 97*(3), 634-654.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics, 36*(3), 385-408.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition, 38*(4), 801-842.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., . . . Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning, 63*(3), 530-566.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*(4), 861-883.

Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 94-112). Bristol: Multilingual Matters.

Loewen, S., & Plonsky, L. (2016). *An A-Z of applied linguistics research methods*. London: Palgrave.

Long, M. H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition, 12*(3), 251-285.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. B. Ginsberg & C. Kramsch (Eds.), *Foreign*

language research in cross-cultural perspective (pp. 39-52). Amsterdam: John Benjamins Publishing.

Long, M. H. (2003). Stabilization and fossilization in interlanguage development. *Language Research, 12*, 335-373.

Long, M. H. (2005). Problems with supposed counter-evidence to the critical period hypothesis. *International Review of Applied Linguistics in Language Teaching, 43*(4), 287-317.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60*(3), 501-533.

MacWhinney, B. (2005). A unified model of language acquisition. In J. F. Kroll, & A. M. B. de Groot (Eds.), *Handbook of bilinguialism: Psycholinguistic approaches* (pp. 49-67). Oxford: Oxford University Press.

MacWhinney, B. (2007). A unified model. In N. C. Ellis, & P. Robinson (Eds.), *Handbook of cognitive linguistics and second language acquisition.* (pp. 341-371). New York: Routledge.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B, 36*(2), 115-128.

Marinova-Todd, S. H. (2003). *Native, near-native or non-native: Comprehensive analysis of the linguistic profiles of highly proficient adult second language learners* (Unpublished doctoral dissertation). Harvard University, Cambridge, Massachusetts.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*(1), 1-71.

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition, 34*(3), 379-413.

Maybery, M., Taylor, M., & O'Brien-Malone, A. (1995). Implicit learning: Sensitive to age but not IQ. *Australian Journal of Psychology*, *47*(1), 8-17.

McDade, H. L., Simpson, M. A., & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *Journal of Speech and Hearing Disorders, 47*, 19-24.

McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied Psycholinguistics, 21*, 395-423.

McLaughlin, B. (1990). The relationship between first and second languages. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp.158-178). Cambridge: Cambridge University Press.

Meara, P. (2005). *LLAMA language aptitude tests: The manual*. Swansea: Lognostics.

Meisel, J. M. (2011). *First and second language acquisition*. Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy, & L. E. Bourne Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339-364). Mahwah, NJ: Lawrence Erlbaum.

Molfese, D. L., & Molfese, V. J. (1997). Discrimination of language skills at five years of age using event-related potentials recorded at birth. *Developmental Neuropsychology, 13*(2), 135-156.

Moyer, A. (1999). Ultimate attainment in L2 phonology. *Studies in Second Language Acquisition, 21*(1), 81-108.

Muñoz, C. (Ed.). (2006). *Age and the rate of foreign language learning*. Clevedon: Multilingual Matters.

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics, 29*(4), 578-596.

Noelle, D. C. (1999). Explicit to whom? Accessibility, representational homogeneity, and dissociable learning mechanisms. *Behavioral and Brain Sciences, 22*(5), 777-778.

O'brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics, 27*(3), 377-402.

O'brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29*(4), 557-581.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Osaka, M., & Osaka, N. (1992). Language-independent working memory as measured by Japanese and English reading span tests. *Bulletin of the Psychonomic Society, 30*(4), 287-289.

Osaka, M., Osaka, N., & Groner, R. (1993). Language-independent working memory: Evidence from German and French reading span tests. *Bulletin of the Psychonomic Society, 31*(2), 117-118.

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publications.

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research, 5*(3), 261-283.

Oyama, S. (1978). The sensitive period and comprehension of speech. *NABE Journal, 3*(1), 25-40.

Paradis, M. (1994). Neurolinguistic aspects of implicit and explicit memory: Implications for bilingualism and SLA. In N. C. Ellis (Ed.), *Implicit and explicit learning of second languages* (pp. 393-419). London: Academic Press.

Paradis, M. (2004). *A neurolinguistic theory of bilingualism*. Amsterdam: John Benjamins Publishing.

Paradis, M. (2009). *Declarative and procedural determinants of second languages*. Amsterdam: John Benjamins Publishing.

Parry, T. S., & Child, J. R. (1990). Preliminary investigation of the relationship between VORD, MLAT and language proficiency. In T. S. Parry, & C. W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 30-66). Englewood Cliffs, NJ: Prentice Hall.

Patkowski, M. S. (1990). Age and accent in a second language: A reply to James Emil Flege. *Applied Linguistics, 11*(1), 73-89.

Payne, J. S., & Whitney, P. J. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *CALICO Journal, 20*(1), 7-32.

Peelle, J. E., Cooke, A., Moore, P., Vesely, L., & Grossman, M. (2007). Syntactic and thematic components of sentence processing in progressive nonfluent aphasia and nonaphasic frontotemporal dementia. *Journal of Neurolinguistics, 20*(6), 482-494.

Pena, E. A., & Slate, E. H. (2014). *Gvlma: Global validation of linear models assumptions.* R package version 1.0.0.2. URL https://CRAN.R-project.org/ package=gvlma

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition, 17*(4), 398-422.

Petersen, C. R., & Al-Haik, A. R. (1976). The development of the defense language aptitude battery (DLAB). *Educational and Psychological Measurement, 36*(2), 369-380.

Pienemann, M. (1984). Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition, 6*(2), 186-214.

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypotheses. *Applied Linguistics, 10*(1), 52-79.

Pimsleur, P. (1966). *Pimsleur language aptitude battery (PLAB)*. San Diego: Harcourt Brace Jovanovich.

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS* (6th ed.) New York: Routledge.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912.

Purcell, E. T., & Suter, R. W. (1980). Predictors of pronunciation. *Language Learning, 30*, 271-287.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 88-94.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning, 63*(3), 595-626.

Rebuschat, P. & Williams, J. N. (2013). Implicit learning in second language acquisition. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 2626-2632). Oxford: Blackwell Publishing.

Renou, J. (2001). An examination of the relationship between metalinguistic awareness and second-language proficiency of adult learners of French. *Language Awareness, 10*(4), 248-267.

Reves, T. (1983). *What makes a good language learner? Personal characteristics contributing to successful language acquisition* (Unpublished doctoral dissertation). The Hebrew University of Jerusalem, Jerusalem, Israel.

Robinson, P. (1995a). Aptitude, awareness, and the fundamental similarity of implicit and explicit second language learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 303-357). Honolulu, HI: University of Hawai'i Press.

Robinson, P. (1995b). Learning simple and complex rules under implicit, incidental rule-search, and instructed conditions. *Studies in Second Language Acquisition, 18*, 27-67.

Robinson, P. (2001). Individual differences, cognitive abilities, aptitude complexes and learning conditions in second language acquisition. *Second Language Research, 17*(4), 368-392.

Robinson, P. (2002a). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211-266). Amsterdam: John Benjamins Publishing.

Robinson, P. (2002b). *Individual differences and instructed language learning*. Amsterdam: John Benjamins Publishing.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics, 25*, 46-73.

Roehr, K. (2006). Metalinguistic knowledge in L2 task performance: A verbal protocol analysis. *Language Awareness, 15*(3), 180-198.

Roehr, K. (2008). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics, 29*(2), 173-199.

Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association, 1*(1), 49-60.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1-36.

Royston, J. P. (1983). Some techniques for assessing multivarate normality based on the Shapiro-Wilk W. *Applied Statistics, 32*(2), 121-133.

Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching, 46*(2), 113-136.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514.

Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 319-353). Cambridge: Cambridge University Press.

Schacter, D. L. (1992). Understanding implicit memory: A cognitive neuroscience approach. *American Psychologist, 47*(4), 559.

Schacter, D. L., Chiu, C. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience, 16*(1), 159-182.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *Consciousness in Second Language Learning, 11*, 237-326.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Scovel, T. (1988). *A time to speak: A psycholinguistic inquiry into the critical period for human speech*. Rowley, MA: Newbury House Rowley.

Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics, 20*, 213-223.

Scovel, T. (1969). Foreign accents, language acquisition, and cerebral dominance. *Language Learning, 19*(3-4), 245-253.

Seliger, H. (1978). Implications of a multiple critical period hypothesis for second language learning. In W.C. Ritchie (Ed.) *Second language research: Issues and implications* (pp. 11-19). New York: Academic Press.

Serafini, E. J., & Sanz, C. (2016). Evidence for the decreasing impact of cognitive ability on second language development as proficiency increases. *Studies in Second Language Acquisition, 38*(4), 607-646.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology, 45*(1), 21-50.

Shanks, D. R. (2003). Attention and awareness in "implicit" sequence learning. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 11-42). Amsterdam: John Benjamins Publishing.

Sharwood Smith, M. (1981). Consciousness-raising and the second language learner. *Applied Linguistics, 11*(2), 159-168.

Simon, J. R., Howard, J. H., & Howard, D. V. (2011). Age differences in implicit learning of probabilistic unstructured sequences. *Journal of Gerontology: Psychological Sciences, 66B*(1), 32-38.

Singleton, D., & Ryan, L. (2004). *Language acquisition: The age factor*. Clevedon: Multilingual Matters.

Skehan, P. (1986). The role of foreign language aptitude in a model of school learning. *Language Testing, 3*(2), 188-221.

Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69-94). Amsterdam: John Benjamins Publishing.

Skehan, P. (2012). Language aptitude. In S. Gass, & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 381-395). New York: Routledge.

Skehan, P. (2015a). Foreign language aptitude and its relationship with grammar: A critical overview. *Applied Linguistics, 36*(3), 367-384.

Skehan, P. (2015b). Working memory and second language performance: A commentary. In Z. Wen, M. B. Mota & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 189-204). Bristol: Multilingual Matters.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. O. Jackson & Y. Yilmaz (Eds.),

Cognitive individual differences in second language processing and acquisition (pp. 17-40). Amsterdam: John Benjamins Publishing.

Skrzypek, A. (2009). Phonological short-term memory and L2 collocational development in adult learners. *EUROSLA Yearbook, 9*(1), 160-184.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59*(2), 205-216.

Sorace, A. (1985). Metalinguistic knowledge and language use in acquisition-poor environments. *Applied Linguistics, 6*(3), 239-254.

Spada, N., Shiu, J. L., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning, 65*(3), 723-751.

Sparks, R., Ganschow, L., Artzer, M., Siebenhar, D., & Plageman, M. (1997). Language anxiety and proficiency in a foreign language. *Perceptual and Motor Skills, 85*(2), 559-562.

St Clair-Thompson, H., & Sykes, S. (2010). Scoring methods and the predictive ability of working memory tasks. *Behavior Research Methods, 42*(4), 969-975.

StatsNZ. (2015). 2013 census QuickStats about culture and identity. Retrieved from http://archive.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-culture-identity/asian.aspx

Suter, R. W. (1976). Predictors of pronunciation accuracy in second language learning. *Language Learning, 26*, 233-253.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics, 38*, 1229-1261.

Suzuki, Y., & DeKeyser, R. M. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65*(4), 860-895.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Upper Saddle River, NJ: Pearson Education.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning, 41*(2), 177-204.

Tolentino, L. C., & Tokowicz, N. (2011). Across languages, space, and time. *Studies in Second Language Acquisition, 33*(1), 91-125.

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? the role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 171-198). Oxford: Oxford University Press.

Ullman, M. T. (2001a). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research, 30*(1), 37-69.

Ullman, M. T. (2001b). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition, 4*(2), 105-122.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*(1), 231-270.

Ullman, M. T. (2005). A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. In C. Sanz (Ed.), *Mind and context in adult second language acquisition* (pp. 141-178). Washington, DC: Georgetown University Press.

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition, 39*(1), 59-95.

VanPatten, B., & Borst, S. (2012a). The roles of explicit information and grammatical sensitivity in processing instruction: Nominative-accusative case marking and word order in German L2. *Foreign Language Annals, 45*(1), 92-109.

VanPatten, B., & Borst, S. (2012b). The roles of explicit information and grammatical sensitivity in the processing of clitic direct object pronouns and word order in Spanish L2. *Hispania,* 270-284.

Wallach, D., & Lebiere, C. (2003). Implicit and explicit learning in a unified architecture of cognition. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 215-252). Amsterdam: John Benjamins Publishing.

Walter, C. (2004). Transfer of reading comprehension skills to L2 is linked to mental representations of text and to L2 working memory. *Applied Linguistics, 25*(3), 315-339.

Wen, Z. (2012). Working memory and second language learning. *International Journal of Applied Linguistics, 22*(1), 1-22.

Wen, Z. (2014). Theorizing and measuring working memory in first and second language research. *Language Teaching, 47*(2), 174-190.

Wen, Z. (2015). Working memory in second language acquisition and processing: The phonological/executive model. In Z. Wen, M. B. Mota & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 41-62). Bristol: Multilingual Matters.

Wen, Z. (2016). *Working memory and second language learning: Towards an integrated approach*. Bristol: Multilingual matters.

Wen, Z., & Skehan, P. (2011). A new perspective on foreign language aptitude research: Building and supporting a case for "working memory as language aptitude". *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, (60), 15-44.

White, L., & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research, 12*(3), 233-265.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition, 27*(2), 269-304.

Williams, J. N. (2009). Implicit learning in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds), *The new handbook of second language acquisition* (pp. 319-353). Bingley: Emerald Group Publishing Ltd.

Williams, J. N. (2012). Working memory and SLA. In S. Gass, & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 427-441). London: Routledge.

Williams, J. N., & Lovatt, P. (2003). Phonological memory and rule learning. *Language Learning, 53*(1), 67-121.

Wong, W. (2001). Modality and attention to meaning and form in the input. *Studies in Second Language Acquisition, 23*(3), 345-368.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*(4), 463-489.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics, 34*(3), 344-368.

Yu, Y., & Wang, B. (2009). A study of language learning strategy use in the context of EFL curriculum and pedagogy reform in china. *Asia Pacific Journal of Education, 29*(4), 457-468.

Zhang, R. (2015). Measuring university-level L2 learners' implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 37*(3), 457-486.

Zhao, Y. (2015). *The effects of explicit and implicit recasts on the acquisition of two grammatical structures and the mediating role of working memory* (Unpublished doctoral dissertation). University of Auckland, Auckland, New Zealand.

APPENDICES

Appendix 1 Erlam's (2009) stimuli of the EI test

1. *New Zealand is greener and more beautiful than other countries.
2. New Zealanders want to keep their country clean and green.
3. Children play rugby well and soccer badly in New Zealand.
4. *People should report the police stolen money.
5. *Everyone loves comic books and read them.
6. The film that everyone likes is Star Wars.
7. People can win a lot of money in a casino.
8. Spending 10 hours in an airplane isn't much fun, is it?
9. People should report a car accident to the police.
10. *People have been using computers since many years.
11. *The software that Bill Gates invented it changed the world.
12. A good teacher makes lessons interesting and cares about students.
13. It is not a good idea for teachers to punish students.
14. *Not everyone can to learn a second language.
15. To speak English well you must study for many months.
16. *It is more harder to learn Japanese than to learn English.
17. Princess Diana loved Prince Charles but divorced him.
18. *If Prince Charles had loved Princess Diana she will be happier.
19. Princess Diana's death shocked the whole world.
20. *The number of Africans with AIDS was increased last year.
21. *The Americans were first to land on the moon, isn't it?
22. If Russia had got to the moon first, America would have been worried.
23. *Everyone wants to know what is President Bush like.
24. *When man invented the motor car, life change for everyone.
25. Last year the population of the world increased a lot.
26. *Young people visit often clubs and drink a lot.

27. *Young women like cigarettes and fast car.
28. *Parents have a responsibility to care for their children.
29. *People worry about their parent health and their children's future.
30. *Every child needs good father.
31. *It is a silly question to ask 'Do a woman need to marry?'
32. *People in love usually want getting married as soon as possible.
33. A wife always wants to know what her husband is doing.
34. It is difficult to ask 'Do you really love me?'

Appendix 2 Stimuli of the WordM test and the TGJT

1 * Kim lives in Hamilton but work in Auckland this year.
2 * My car is more faster and more powerful than your car.
3 * This building is more bigger than the house over there.
4 David works very hard but earns very little.
5 * Kim went to the school to speak to her children teacher two days ago.
6 Peter has won a prize for the book he wrote.
7 Kim wants to buy a computer this weekend.
8 Mary is taller than her older sisters now.
9 * David is still living in his rich uncle house in the suburb.
10 * Hiroshi live with his friends Koji and Takeshi.
11 Something bad happened last weekend.
12 Hiroshi found some keys on the ground.
13 Did Cathy cook dinner last night?
14 They did not come at the right time unfortunately.
15 David stayed at home all day and finished the book.
16 They enjoyed the party very much.
17 * Kim bought two present for his friend Jim's children.
18 * The bird that my brother caught it has died all of a sudden.

19 * Did Keiko completed her mathematics homework yesterday?

20 Does Kim live close to the university?

21 * The boys went to bed late last night, is it?

22 Japan is a very interesting country to visit.

23 * He reported his father the bad news.

24 Hiroshi received a letter from his father yesterday.

25 * I can to speak French very well.

26 Joseph flew to Washington to meet the President's advisor.

27 * The boat that my father bought it has sunk in the storm.

28 * He has been living in New Zealand since three years.

29 Keiko eats a lot of fish every week.

30 Rosemary reported to the police the crime.

31 Bill wanted to know where I had been.

32 * If he had been richer, she will marry him.

33 I must finish my homework tonight.

34 I haven't seen him for a long time.

35 * She likes always watching television at home.

36 * I must to brush my teeth now.

37 * David says he wants buying a car next week.

38 * She wanted to know why had he studied German at University.

39 * Tom wanted to know what had I learned in my dancing class.

40 Keiko has been studying in Auckland for three years.

41 She is working very hard, isn't she?

42 * The population of New Zealand was increased between 1990 and 2000.

43 * If he hadn't come to New Zealand, he will stay in Japan.

44 Pam wanted to know what I had told John and Jerry.

45 * The teacher explained John the answer in detail.

46 He plays soccer very well.

47 * An accident was happened on the motorway yesterday.

48 * I saw very funny movie last night.

49 If she had worked hard, she would have passed the exam this semester.

50 I can cook Chinese food very well.

51 The house that they have rented has a nice view.

52 * Joseph wants finding a new job next month.

53 * We will leave tomorrow, isn't it?

54 If he had bought a ticket, he might have won the best prize.

55 * She writes very well stories in English.

56 Her English vocabulary increased a lot last year.

57 The teacher explained the problem to the students in the class.

58 Martin says he wants to get married next year.

59 * I have been studying English since a very long time.

60 * Martin sold a few old coins and stamp to a shop near his home.

61 * Did Martin visited his father yesterday?

62 That book isn't very interesting, is it?

63 David left some pens and pencils at school yesterday.

64 I think that he is nicer and more intelligent than all the other students.

65 Keiko spoke to the professor's secretary last week.

66 *They had the very good time at the party.

67 * Joseph miss an interesting party last weekend.

68 * Martin completed his assignment and print it out two days ago.

Appendix 3 Davies's (1991) cloze test

1. Here is _____ book we were talking about.
2. If_____ of you brings a chair, you will _____ be able to sit down.
3. I saw _____ rain all the time I was in Britain.
4. He would have to ask his wife's permission before lending the case because it was _____ his _____ hers.

5. When you have finished _____ book, go on to _____ one.

6. Some men cultivate their farms _____ _____ work in factories.

7. _____ John _____ you is wrong; you can't _____ be right.

8. At the end of the visit, because the children were hungry and tired we _____ _____ make them wash their hands before eating.

9. This school has no more places _____ for boys _____ for girls.

10. We can see that _____ buses _____ cars have been along this road.

11. As I am a little deaf, I find that I can _____ always hear what is said _____ the telephone.

12. How _____ I hold the baby so that he _____ not cry?

13. I should like the light on now, please _____ _____ wait until dark.

14. Please take the bandages off my head, _____feels sorer and _____.

15. I haven't used _____ library at all _____ it has been open every day.

16. We shall certainly go for a picnic tomorrow _____ you come _____ not.

17. Some men enjoy football while _____ _____ hate the game.

18. The weather, _____ _____ the floods, has improved this year.

19. For breakfast I always eat porridge _____ eggs _____ _____ bacon.

20. _____ you hear him say that he will _____ _____visit the tax-office today?

Appendix 4 Answers to the cloze test

No.	Davies (1991)	Other acceptable answers
1	that; the	
2	each, all; one, then	one; all, both, probably
3	no	it, much, this, that

4	not, but	both, and
5	that, this	one, another
6	while others	and others, while some
7	either, or, both	
8	did not	could not
9	vacant, than; either, or	suitable, or
10	both, and	no, or
11	not, over; not, on	
12	should; will	do, can, might; does, will
13	do not	lets not
14	it, sorer; it, aches	it, heavy/hurts
15	the, (al)though	that, since
16	whether, or	if, or
17	others, really; other, men	some women
18	apart from; despite all	expect for; ever since
19	or, but, never; and, but, not	and, with no/some
20	did, try, to	did, most certainly/likely

Appendix 5 Annotated Chinese MKT

一、本题共 17 小题，题干中下划线部分均有语法错误。请从 a, b, c, d 四个选项中选出能够解释题干中的语法错误的最佳答案。

例 1：

Keiko said, ‘ I have lost mine ring’.

a. Replace the word ‘mine’ with ‘my’.
b. Mine cannot be used as a possessive (所有格) word.
c. Should be ‘her ring’ because Keiko is the subject (主语).
d. Before a noun use the possessive (所有格) adjective, not the pronoun (代词).

正确答案: d

例 2:

He saw a elephant.

a. The word ‘elephant’ refers to the normal verb (动词).

b. We must use ‘elephant’ instead of ‘a elephant’.
c. You should use ‘an’ not ‘a’ because elephant starts with a vowel sound (元音).
d. The wrong form of indefinite article (不定冠词) has been used.

正确答案: c

Now Start

1. You must to wash your hands before eating.
   a. ‘Must to’ is the wrong form of the imperative (祁使语气).
   b. Change to ‘must have to wash’ to express obligation.
   c. Modal verbs (情态动词) should never be followed by a preposition (介词).
   d. After ‘must’ use the base form (原形) of the verb not the infinitive (动词不定式).

2. Hiroshi wants visiting the United States this year.
   a. ‘Visiting’ should be written in the base form (动词原形).
   b. The verb following ‘want’ must be an infinitive (动词不定式).
   c. We cannot have two verbs (动词) together in a sentence.
   d. It should be ‘visit’ because the event is in the future.

3. Martin work in a car factory.
   a. Work is a noun (名词) so it cannot have the subject (主语) ‘Martin’.
   b. We must use the present simple tense (一般现在时) after a pronoun (代词).
   c. We need ‘s’ after the verb to indicate third person (第三人称) plural (复数).
   d. In the third person (第三人称) singular (单数) the present tense (现在时) verb takes ‘s’.

4. If Jane had asked me, I would give her some money.
   a. ‘would’ is conditional (条件句) so it should appear in the ‘if’ clause (if从句) not the main clause (主句).
   b. The first clause (小句) tells us that this is an impossible condition, so use the subjective.
   c. We must use ‘would have given’ to indicate that the event has already happened.
   d. When ‘if’ clause is in the past perfect tense (过去完成时), main clause (主句) verb is in the past conditional.

5. Learning a language is more easier when you are young.
   a. ‘More’ is an adjective (形容词) so we must use ‘easily’ not ‘easier’.
   b. The comparative (比较级) ending of a two-syllable (双音节) adjective is ‘er’.

c. The ‘er’ ending indicates comparison, so ‘more’ is not needed.
d. You cannot have two adjectives together in the same sentence.

6. Keiko grew some rose in her garden.
   a. The noun is countable (可数), so after ‘some’ use the plural form (复数).
   b. The wrong adjective (形容词) has been used before ‘rose’.
   c. A noun must always have ‘a’ or ‘the’ before it.
   d. Use ‘a few’ not ‘some’ with countable nouns.

7. His school grades were improved last year.
   a. The verb ‘improve’ can never be used in the passive form (被动语态).
   b. We should insert ‘by him’ after the verb to indicate the agent (施事者).
   c. Use ‘improved’ as the sentence refers to a specific event last year.
   d. ‘Improve’ should take the active form (主动语态) even though the subject (主语) is not the agent (施事者).

8. Martin lost his friend book.
   a. We need possessive ‘s’ (所有格) to show that the friend owns the book.
   b. You cannot have two nouns (名词) next to one another in a sentence.
   c. The verb refers to a personal object, so must have an apostrophe (撇号).
   d. Insert ‘of’ before book to show that it belongs to the friend.

9. Keum happen to meet an old friend yesterday.
   a. It took place yesterday, so use a past tense (过去时) verb ending.
   b. Third person singular (第三人称单数) verbs always have an ‘s’ ending.
   c. We don’t use a preposition (介词) after the verb (动词) ‘happen’.
   d. ‘Happen’ never follows the subject (主语) of a sentence.

10. Because he was late, he called taxi.
    a. Insert ‘a’ before taxi because it is not a specific one.
    b. Use ‘some taxis’ because taxi cannot be singular (单数).
    c. We must always use ‘the’ before countable nouns (可数名词).
    d. Use the indefinite article (不定冠词) because the taxi is unique.

11. They were interested in what was I doing.
    a. In embedded questions (嵌入式疑问句) the word order (词序) is the same as that in statements (陈述句).
    b. Change the word order, because ‘what’ is always followed by a pronoun (代词).
    c. The subject (主语) should always come in front of the verb (动词) after question words.
    d. The clause (小句) ‘What was I doing’ should be followed by a question mark.

12. Does Liao has a Chinese wife?
    a. With questions, always use the auxiliary (助动词) ‘have’.
    b. We must use the base form (原形) after ‘do/does’.
    c. Use ‘have’ not ‘has’ because ‘does’ is in the past tense (过去时).
    d. The word order changes when we use the question form.

13. Jenny likes very much her new job.
    a. Adverbial phrases (副词短语) should occur after nouns not verbs.
    b. An adverb (副词) should not come between a verb and its object (主语).
    c. The phrase ‘very much’ always occurs at the end of a sentence.
    d. The adverbial phrase must always precede the verb.

14. They have already finished, isn’t it?
    a. We cannot use ‘it’ because the main verb ‘finish’ does not have an object (宾语).
    b. ‘have’ should be used instead of ‘is’ in all question tags (反意疑问句句尾) referring to past time.
    c. The tag question (反意疑问句) should be positive because the main verb is in the affirmative (肯定).
    d. The form of the question tag (反意疑问句句尾) must relate to the subject and verb in the main clause.

15. He has been saving money since 10 years.
    a. The wrong conjunction (连词) has been used in the time clause.
    b. We cannot use ‘since’ because the exact date is specified.
    c. Use ‘for’ following any verb in the past perfect continuous tense (过去完成进行时).
    d. Use ‘for’ not ‘since’ for a noun phrase referring to a period of time.

16. I explained my friend the rules of the game.
    a. The indirect object (间接宾语) must never precede the direct object (直接宾语) of a verb.
    b. ‘Explain’ (unlike the verbs ‘tell’ and ‘give’) can only have one object.
    c. After ‘explain’ we must insert a preposition (介词) before the indirect object.
    d. The preposition ‘to’ is always used for the dative form (与格) of a noun or pronoun.

17. The cake that you baked it tastes very nice.
    a. Omit ‘that’ when the relative pronoun (关系代词) is subject of the clause.
    b. We should use ‘which’ instead of ‘that’ when referring to things.

c. Omit ‘it’ in the relative clause (关系从句) because it refers to same thing as ‘that’.
d. Omit ‘that’ when using ‘it’ in the relative clause to avoid having two pronouns.

二、 请阅读以下段落，在该段落中找出一个与表格对应的语法成分并写在空格内。同一答案可用于多个问题。

The materials are delivered to the factory by a supplier, who usually has no technical knowledge, but who happens to have the right contacts. We would normally expect the materials to arrive within three days, but this time it has taken longer.

Question No.	*Grammatical feature*	*Example*
例	definite article 定冠词	the
1	verb 动词	
2	noun 名词	
3	preposition 介词	
4	passive verb 动词被动式	
5	adjective 形容词	
6	adverb 副词	
7	countable noun 可数名词	
8	indefinite article 不定冠词	
9	relative pronoun 关系代词	
10	auxiliary verb 助动词	
11	modal verb 情态动词	
12	past participle 过去分词	
13	conjunction 连词	
14	finite verb 限定性动词	
15	infinite verb 非限定性动词	

16	agent 施事者	
17	comparative form 比较级	
18	pronoun 代词	

三、在下列句子中，请用下划线标出括号内的句子成分。

1. Poor little Joe stood out in the snow. (Subject 主语)
2. Joe had nowhere to stay. (Infinitive 动词不定式)
3. The policeman chased Joe down the street. (Direct object 直接宾语)
4. The woman gave him some money. (Indirect object 间接宾语)

Appendix 6 Zhao's (2015) stimuli of the NonWord test

No. of Items	Content of Items
1.	mou1, shuan2
2.	te3, chuo2
3.	dei4, jiong1
4.	rui1, gei4, nen3
5.	jiong1, qun4, niao1
6.	diu4, zhun2, xiong3
7.	nie2, kei4, run3, zang2
8.	lia4, miu3, le2, neng4
9.	niao1, mie3, kao1, sai3
10.	die3, le2, dei4, ruan1, miu3
11.	gei4, fou1, ruan1, xiong3, kao1
12.	de4, nue3, ken4, min1, mang1
13.	teng3, lue3, fo4, dong2, fou1, diu4
14.	nen3, yue2, run3, fo4, lue3, pou2
15.	dian2, teng3, suan2, pou2, ran1, qiong4
16.	mie3, rong1, diu4, chuo2, ka2, dei4, niang3
17.	ran1, shuan2, qiong4, mang1, sai3, ken4, zhun2

18.	ruan1, lue3, mou1, zhun2, gei4, zang2, yue2
19.	nen3, nuo1, teng3, ka2, fo4, rui1, miu3, run3
20.	qun4, dian2, de4, rui1, dong2, niang3, nue3, nuo1
21.	kei4, ran1, fou1, neng4, die3, chuo2, lia4, nie2
22.	yue2, suan2, nie2, kei4, zang2, mie3, rong1, nin4, ka2
23.	ken4, kao1, nin4, nue3, rong1, le2, te3, qiong4, niao1
24.	shuan2, pou2, jiong1, suan2, sai3, qun4, de4, min1, mang1

Note. The numbers following *pinyins* represent Chinese lexical tones.

Appendix 7 Zhao's (2015) stimuli of the SSpan test

测试一	我们待人应心胸宽广，对已要严格要求。
	*经过漫长的急救，小狗总算救活了兽医。
	*我国天然货物很多，种类只有煤和石油。
测试二	*人人都买手机，因为它用起来很不方便。
	这个节目非常有趣，适合全家老少观赏。
	*我昨天工作到很晚，直到回家才十二点。
	*这么重要的事给他做，我真的不能小心。
	*大地震过后，各国都主动地捐款来购买。
测试三	*他不仅夸奖顶头上司，还给上司加薪水。
	写论文引用资料时，应注明文献的出处。
	*我出门没带钱，不幸遇到老友才没丢脸。
	美国东北部出现罕见的暴雪，损失惨重。
	老先生医术精湛，给患者留下深刻印象。
	如果不经常吸收新知识，很容易被淘汰。
测试四	*我们必须认真修正五香十色的财经法规。
	*这项工程已经竣工，可是无法准时完成。
	受国际油价飙升影响，物价呈上涨趋势。
	当义工不仅让自己充实，也使别人受惠。

测试五	*老张疑心太重，不论人家说什么他都信。 *这次我考得不理想，因为时间不够充满。 外面风声很紧，这件东西根本脱不了手。
测试六	*公路暂时封闭，导致飞机不能正常起降。 *他每天一下课就从电影院跑出来看电影。 这位候选人的政治见解很好，深得民心。 *我每次主动打扫卫生，妈妈都会训斥我。 *他显然从错误中吸取了教训，又犯了错。 *这列火车会到过那站，但是不会停下来。
测试七	爸爸常对我说，一分耕耘就有一分收获。 *每到清明，总有五花八门的人采购礼物。 由于生活压力太大，他的睡眠质量不好。 他坚持每天读书，不断吸收并积累知识。
测试八	专柜售货员说这双鞋看起来非常适合你。 恐怖事件发生后，各国都加强防范措施。 *看到热腾腾的鸡汤，无法令人五颜六色。 你千万不要错过这个千载难逢的好机会。
测试九	最近天气多变，你们出门要多加件外套。 *他学习用功，顺利地没有通过这次考试。 网络犯罪猖獗，应该采取有效防范措施。
测试十	*这事情与我无关，对不起，都是我的错。 *学校一定会对大家优异的表现得到批评。 三餐饮食要均衡，毕竟药补还不如食补。 *这些贝壳各有特色，都不值得慢慢欣赏。 哥哥做事认真负责，领导对他信任有加。
测试十一	*只要有一点瑕疵，商品都可以退换顾客。 考试快到了，他正为考试做最后的冲刺。 *虽然他有过失，你也犯不着当众表扬他。 在高考失败之后，他就变得失魂落魄了。

	*他很愤世嫉俗，从来都不会抱怨和挑剔。
	研究表明，饮食习惯与肥胖有一定关系。
测试十二	经济不景气，想找到好工作非常不容易。
	*姐姐打电话，不论一高兴，就忘了时间。
	为了爸爸身体健康，全家人都劝他戒酒。
	因为事先没有规划，今天才会如此失败。
	*疫情日趋严重，使得各产业恢复了正常。
测试十三	*流感病例越来越多，大家不必再量体温。
	抱怨与诉苦，对减轻痛苦一点用都没有。
	哥哥没准备好就去比赛，他感到很紧张。
	你桌球打得最好，应代表班级参加比赛。
测试十四	我们要从长远的角度，来思考两岸关系。
	*我打工赔了钱，不再需要跟父母要钱花。
	我在德国留学的那段日子，常想起家人。
	黄金枫开始落叶的时候，景色美不胜收。
	关节内视镜手术可治疗肩颈关节的损伤。
测试十五	*学校放假期间，学生每天都按时去上课。
	应该趁年轻多读一些书，不要虚度时光。
	*他看书不多，要是历史类，就是地理类。
测试十六	*他不敢回家，父母不知该怎么向他交代。
	*开会时我劝你闭上眼睛，认真看着老板。
	人口老龄化已成为发达国家最大的隐患。
	虽然她现在已不发烧了，但仍有些虚弱。
	*这是全球电脑市场增长最快的地区之间。
	*他游得实在太好了，平常应该多练一练。

Appendix 8 Questionnaire

English Learning Experience Questionnaire

Name: _____ Email: _____

Tel: _____

Instructions:

In this questionnaire, the researcher is interested in your English learning experience in China and New Zealand, as well as your current English usage situation. You would want to RECALL, as much as possible, your learning experiences in primary school, secondary school, and maybe also in university. If you can NOT remember clearly, please put N/A as an answer.

The information you provided will be held confidential, please follow the instructions and complete ALL the questions.

Part 1. Learning English in China：Before university

1. How old were you when you started to learn English at school in China?

   _____ years old.

2. How many years did you spend in learning English in China?

   ___ years in primary school;

   ___ years in middle school;

   ___ years in high school;

3. How many English classes (40 to 50 minutes/class) did you have per week in China?

   ___ in primary school;

   ___ in middle school;

   ___ in high school;

4. When you were at middle/high school in China, how many students were in your class?

   A. Less then 30.

   B. 30 to 40.

   C. 40 to 50.

   D. More than 50.

5. What language did your English teacher in middle school mainly use in class?
    A. More English.
    B. More Chinese.
    C. A mixture of both.
6. What language did your English teacher in high school mainly use in class?
    A. More English.
    B. More Chinese.
    C. A mixture of both.
7. Which of the following would better describe your way of learning English at school (i.e. before university) in China?
    A. Text book based: a lot of time was spent on studying text book and grammar rules.
    B. Communication based: a lot of time was spent on using English to express my own ideas.
8. Which of the following would better describe your way of learning grammar at school in China?
    A. The teacher explained the grammar rules first, and then asked me to memorize.
    B. The teacher gave sample sentences and asked me to find out the grammar rules by myself.
9. Which of the following would better describe your way of learning vocabulary at school in China?
    A. The teacher taught vocabulary first and then asked me to recite and dictate.
    B. The teacher asked me to infer the meaning of unknown new words in reading.
10. Which of the following would better describe your way of learning speaking at school in China?

A. I learned speaking by reading text book passages after the teacher in class.

B. I learned speaking by using English to communicate with classmates.

11. Which of the following would better describe your way of learning listening at school in China?

A. I learned listening by doing text book based listening exercise.

B. I learned listening from other resources, such as news, movies, songs provided by my teacher.

12. Did you attend extra-curriculum English classes (e.g. private English schools and training courses) when you were in China?

A. Yes. (Continue to question 13 and 14)

B. No. (Skip question 13 and 14)

13. How many hours did you spend per week at extra-curriculum English classes?

_____hours.

14. Which of the following would better describe your learning at extra-curriculum English classes?

A. Most of the time was spent on enhancing the grammar and vocabulary knowledge.

B. Most of the time was spent on using English to engage in conversations.

15. Did you ever learn from a native English speaker (a foreign teacher) when you were in China?

A. Yes. (Continue to question 16 and 17)

B. No. (Skip question 16 and 17)

16. How many hours did you spend per week with your foreign teacher, and for how many years/months?

_____hours per week, for ____years/months.

17. What did you learn from your foreign teacher?

A. I learned more about English grammar.

B. I learned how to use English to express my ideas.

Part 2. Learning English in China: in university

18. Did you major in English at university?

A. Yes. (Answer questions 19 to 22)

B. No. (ONLY answer questions 19, 20, and 21)

19. How many English classes did you have per week (50min to 1h per class)?

_____

20. How many students were in your class when you were in university?

A. Less than 30.

B. 30 to 40.

C. More than 40.

21. Did you have a foreign teacher (i.e. a native speaker of English) when you were in university?

A. Yes, I had classes with my foreign teacher for _____ hours a week.

B. No.

22. Did you take part in TEM 4 and TEM 8 examinations when you were in university?

A. Yes, I passed TEM 4 and TEM 8.

B. Yes, I passed TEM 4.

C. No.

Part 3. Learning English in New Zealand

23. How old were you when you moved to New Zealand?

_____ years old.

24. When you first came to New Zealand, were you able to talk fluently to Kiwis?

A. Yes, I had no problem in communication at all.

B. Yes, no big problems but sometimes difficult.

C. No, I took a while to be able to speak in English.

25. When you first came to New Zealand, were you able to read and write in English when necessary (e.g. read instructions, fill out documents)?

A. Yes, I had no problem with that.

B. Yes, no big problems but sometimes difficult.

C. No, I took a while to be able to read and write in English.

26. Did you study at language schools/ language courses after arriving in New Zealand?

A. Yes, for ____weeks/months. (Continue to question 27 and 28)

B. No. (Skip question 27 and 28)

27. Which of the following would better describe your learning at language schools/courses in New Zealand?

A. I learned more about English grammar.

B. I was given opportunities to practice using English to communicate.

28. How many students were in your class when you studied at language schools/courses in New Zealand?

A. Less than 20.

B. 20-30.

C. 30-40.

D. More than 40.

29. Did you study in New Zealand secondary schools?

A. Yes, for ____ years.

B. No.

30. Did you study in New Zealand universities (including postgraduate study)?

A. Yes, for ____ years.

B. No.

Part 4. Current status: employment and language usage

31. How old are you now?

_____ years old.

32. Did you work in an environment where English is the working language?

A. Yes, for _____ years.

B. No.

33. Are you now working in an environment where English is the working language?

A. Yes.

B. No.

34. Altogether, how many years have you spent living in a country where English is widely spoken (including New Zealand)?

_____ years.

35. Do you use computers at work or for entertainment at home?

A. Yes, I use computers quite often.

B. No, I don't use computers a lot.

36. Which of the following apply to you? (choose ALL that apply to you)

A. I go to an English-speaking church regularly.

B. I socialize with my English-speaking friends regularly.

C. I mainly socialize with other Chinese people.

D. I speak more Chinese than English with my family at home in New Zealand.

E. I speak more English than Chinese with my family at home in New Zealand.

F. I speak English with my children at home in New Zealand.

G. I speak Chinese with my children at home New Zealand.

H. I speak English and Chinese with my children at home in New Zealand.

I. I like to attend local activities and talk to Kiwis in English.

J. I often watch New Zealand TV or listen to local radio programmes in English.

K. I often visit New Zealand news websites or read English newspaper (e.g. NZ herald).

L. I often watch English TV shows/ movies, including those from US, UK, etc.