ABSTRACT

Title of Dissertation:	TEMPORAL DISTRIBUTION OF PRACTICE AND INDIVIDUAL DIFFERENCES IN THE ACQUISITION AND RETENTION OF L2 MANDARIN TONAL WORD PRODUCTION
	Man Li, Doctor of Philosophy, 2017
Dissertation directed by:	Dr. Robert M. DeKeyser, Professor, Second Language Acquisition

This dissertation investigated the effects of temporal distribution of practice (relatively massed vs. distributed) on the learning and retention of oral Mandarin tonal word production by native English-speaking adults within the theoretical framework of skill acquisition and retention theories. The present study focused on oral production of Mandarin two-syllable words as a function of temporal distribution of practice. It also explored whether the effects of this distribution differ depending on the type of knowledge to be acquired or retained (declarative word knowledge vs. skills in oral production) and on individual differences in cognitive aptitudes (including working memory, phonological short-term memory, declarative memory, procedural memory, and musical aptitude).

Eighty native English-speaking adults who did not have any prior knowledge of a tonal language completed all sessions of the study and provided data for analysis.

These participants were randomly assigned to four experimental conditions, i.e., Condition A with a 1-day ISI (intersession interval) and a 1-week RI (retention interval), Condition B with a 1-day ISI and a 4-week RI, Condition C with a 1-week ISI and a 1-week RI, and Condition D with a 1-week ISI and a 4-week RI. Each participant came in for five sessions. All participants completed a set of cognitive aptitude tests and underwent the same number and content of training sessions, which differed only on training or testing schedules.

The results showed that the effects of ISI and RI differed depending on the type of knowledge/skill to be retained, declarative versus procedural. For the retention of declarative knowledge, RI had a robust effect: the longer the RI, the worse the retention. Spacing, or distributed practice seemed to improve long-term retention of declarative knowledge; however, this ISI effect was much weaker. With regard to procedural knowledge retention, ISI seems to play a role, but not RI, and it was massed practice that had an advantage over distributed practice. Musical aptitude, working memory, and declarative memory ability were found to play facilitative roles in L2 learning of Mandarin tonal word productions. Procedural memory ability was found to interact with ISI and RI for various RT outcome measures.

TEMPORAL DISTRIBUTION OF PRACTICE AND INDIVIDUAL DIFFERENCES IN THE ACQUISITION AND RETENTION OF L2 MANDARIN TONAL WORD PRODUCTION

by

Man Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Dr. Robert M. DeKeyser, Chair
Dr. Nan Jiang
Dr. Steven J. Ross
Dr. Jared A. Linck
Dr. Min Wang

Acknowledgements

I would like to express my heartfelt gratitude to my advisor, Dr. Robert DeKeyser, for his guidance, unwavering support and warm encouragements throughout this whole process. Thank you Robert for always being available and dependable whenever I had questions or needed help during each stage of this research; without you, this dissertation would not have progressed so smoothly.

I would also like to thank my committee members, Dr. Nan Jiang, Dr. Steven Ross, Dr. Jared Linck, and Dr. Min Wang for their support of my dissertation work. I'm deeply indebted to Dr. Nan Jiang who has been a mentor throughout my doctoral study here and for being a skeptical and helpful committee member. I'm also very grateful to Dr. Steven Ross for his guidance and insightful suggestions on statistical analyses for this project as well as earlier projects I conducted here. My deep gratitude also goes to Dr. Jared Linck, whose insightful suggestions and constructive comments turned out to be extremely helpful and useful to me. My gratitude also goes to Dr. Min Wang who served as a committee member for my second qualifying paper and for being the Dean's representative on my dissertation committee.

I'm also grateful to the individuals who shared their aptitude tests with me for my dissertation research. My thanks go to Dr. Robert Slevc who shared the musical ability tests with me, Dr. Michael Dougherty who provided the Shapebuilder and delivered the data to me, Drs. Anita Bowles and Valerie Karuzis for sharing the Pitch STM test, Dr. Katherine Martin for sharing the non-word repetition task, Dr. Scott Kaufman for sharing the SRT task, and Dr. DeKeyser for purchasing the CVMT test for me.

I'd also like to thank my two research assistants in Spring 2017, Lia Kaufman and Gavri Schreiber, for their help with scoring two of the aptitude tasks. I'm also grateful to all my participants who completed all five sessions strictly following the specific schedules during their busy semesters. Thanks also go to the individuals who helped doing audio recordings or scoring the outcome task for me. Special thanks go to my friend Jianhui Zhou who made time to listen to me about my design and advised me on conducting statistical analyses.

I would also like to thank the individuals whose comments at the early stage helped shape my dissertation project, including Drs. Mike Long, Nan Jiang, Kira Gor and all my fellow SLAers when I ran my ideas at 649R in Spring 2016, and Professors Mei Kong and Jungjung Lee-Heitz in the Chinese program who made time to discuss and share with me the Chinese linguistic structures that they found hard for English native speakers to learn. I'm also grateful to Professors Minglang Zhou, Guiling Hu and Yuli Wang for their mentorship when I worked as a TA in the Chinese program. The teaching experience there further inspired my research interest in studying L2 learning of Mandarin Chinese.

I'd also like to thank the *Language Learning* dissertation grant and the SLA research funds that provided the necessary financial support for the extensive data collection for this research. I'm also grateful to the Ann G. Wylie dissertation fellowship from the Graduate School that made it possible for me to work full time on dissertation writing this semester. I'd also like to thank the SLLC and the Ph.D. program in SLA for the financial support for my first four years here at UMD.

I'm grateful to my peers too, for the journey we went through together in the SLA program, in particular, Eric Pelzl for all the discussion we had about Mandarin tone learning and instruction which contributed directly to my dissertation research, Yuichi Suzuki for always being so inspiring, Qian Zhou, Kyoko Hillman, Gisela Granena, Susan Benson, Jiyong Lee, Assma and Buthainah Thowaini, Payman Vafee, Sunhee Kim, Eunsoo Kang, Wei Yi, and Qi Zheng.

I'm also grateful to the SLLC administration community who provided help when I needed in conducting all my research projects in Jimenez and during my study here at UMD, especially Pamala Deane, Dr. Lauretta Clough, David Watson, Michelle Dove, Josiland Chambers, Yuk Fan Tai, Nicco Cooper, Claire Goebeler, and Jeffrey Maurer.

On my academic journey towards my doctoral degree, I'm also grateful to the friends I met at GSU, in particular, Jack Hardy, Yanbin Lu, Liang Guo and Weiwei Yang, Pam Pearson, and Caroline Payant. I'm also grateful to the professors and collaborators I met there, especially Dr. Diane Belcher, Dr. Sara Weigle, Dr. Eric Friginal, and Dr. Viviana Cortes. My gratitude also goes to my M.A. advisor Dr. Xiaoqing Qin who introduced me to the wonderful field of SLA.

Finally, my deepest gratitude to my family. To my husband Fuxin, for his understanding, support, encouragements, and help whenever I needed it; to my parents and parents-in-law for their generous help when we most needed it; to Uncle Cai and Aunt Yan for encouraging me to study abroad and pursue my academic goals, and a lot more; and to my daughter who gave me so much joy, love, and understanding during this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

An important goal of research in instructed second language acquisition (SLA) has been to find ways to optimize instructional interventions to facilitate L2 learning. Over the past four decades, considerable attention has been devoted to investigating the comparative effectiveness of explicit versus implicit instruction (see Norris & Ortega, 2000; Goo, Granena, Yilmaz, & Novella, 2015, for two meta-analyses), in particular, of different types of corrective feedback, such as recasts, metalinguistic feedback, explicit correction, and prompts (see S. Li, 2010; Lyster & Saito, 2010, for two meta-analyses), and of comprehension- versus production- based instruction (see DeKeyser & Prieto Botana, 2015, for a narrative review; Shintani, 2015; Shintani, Li, & Ellis, 2013, for two meta-analyses) on the development of L2 grammatical knowledge. In addition to the overall effectiveness of different types of treatment, which serves to answer the question of *whether* and *to what extent* a treatment works, the field has seen an increasing interest in exploring the interactions of treatment type with individual differences in learners, i.e., ATI research (see Roehr, 2013; Vatz, Tare, Jackson, & Doughty, 2013 for reviews) and with the target language features/structures (see Spada & Tomita, 2010, for a meta-analysis), for the purposes of understanding *how* and *why* it works (i.e., how and why a treatment works for some learners but not the others, and with certain linguistic structures but not others) and further theorizing about the nature of the underlying learning processes or mechanisms (see DeKeyser, 2012).

Up to this point, however, the field of instructed SLA has paid much less attention to the effects of a higher-order contextual variable, i.e., temporal distribution

of instruction/practice, on L2 learning, despite repeated calls for more research in this area (e.g., DeKeyser, 2015; Ellis, 2006). Given that L2 learners, especially those in a foreign language (FL) context, have limited time devoted to FL learning, and that achieving high-level proficiency in using the L2 takes a huge amount of practice over time, a practically important question for FL learners and educators is how to distribute study/practice time in order to optimize learning and retention. In addition to practical pedagogical utilities, research into the temporal distribution of practice on second language acquisition and retention and its interactions with cognitive variables and/or target structure characteristics also have potential for theoretical contribution to a better understanding of the underlying mechanisms of cognitive skill acquisition and retention in general and of L2 learning and retention in particular.

The issue of temporal distribution of practice has been extensively studied in cognitive psychology for its value in contributing to a better understanding of human memory and cognition. This issue has also attracted substantial interest from researchers in educational psychology and motor skills learning, due to its potential for practical utility. In SLA, the effects of temporal distribution of instruction or practice have been investigated in macro-level program evaluations of the learning of global L2 skills, such as reading, listening, writing and speaking performance (see Serrano, 2011, for a review). The effects of temporal distribution of practice have also been investigated with focus on discrete target items, such as vocabulary and grammar. For L2 vocabulary learning, previous studies have only focused on the memorization of paired associates in the visual domain (e.g., Bloom & Shuell, 1981; Nakata, 2015; Pavlik & Anderson, 2005; Schuetze, 2015), without taking oral word

production as part of word learning. As for L2 grammar learning, only a few recent studies have examined this issue (Bird, 2010; Miles, 2014; Rogers, 2015; Suzuki, 2017a, 2017b, Suzuki & DeKeyser, 2015, 2016). To the best of my knowledge, no studies have examined this temporal issue in the context of word learning that takes auditory perception, oral production, and L2 phonology into account.

The present dissertation attempts to examine the effects of temporal distribution of practice (relatively more massed vs. distributed) on the automatization and retention of L2 Mandarin word production by a group of naïve native speakers (NSs) of English. Oral word production from meaning to sound is a complex task that is both cognitive and motor, because it involves both cognitive memorization of meaning-word mappings and speech motor articulation. Mandarin Chinese is a tonal language that employs pitch variations to distinguish lexical meaning (Chao, 1948). Speakers of a non-tonal language, such as English NSs, need to learn to attend to this new phonological feature (i.e., lexical tones) in learning each Mandarin word. In addition, the present study focuses on the learning of oral production of *disyllabic* words. As disyllabic words may involve tonal changes, this adds another dimension of learning difficulty; learners need to learn the tonal change rule and apply the rule in the appropriate context for oral production. Therefore, learning to orally produce disyllabic Mandarin words is a rather complex task for English NSs, which involves the learning of both declarative knowledge about meaning-word mappings and Mandarin tones, and procedural knowledge for the oral production of the words.

As the task of learning oral Mandarin word production involves the learning of a mix of different types of knowledge, declarative and procedural, a question arises

as to how to structure practice time for optimal learning and retention of this skill. Would it be more effective if learners practiced the sub-components of this complex skill at different temporal schedules? To answer this, we need to first examine whether the effects of temporal distribution of practice differ by the type of knowledge (declarative vs. procedural) to be acquired and retained. In addition, from a skill acquisition theory perspective, the acquisition of a skill typically goes through three stages from declarative stage to proceduralization and automatization; the question arises as to whether the effects of the temporal distribution of practice are the same for all stages of learning. Would less spacing, or more intensive practice, be more effective for the proceduralization stage and for incipient automatization, considering that less spacing would provide more available access to and retrieval of the declarative knowledge that is critical for proceduralization (DeKeyser, 2007b), and that procedural knowledge is more robust and much less vulnerable to memory decay than declarative knowledge (Kim, Ritter, & Koubek, 2013)? These are the questions the present dissertation attempts to explore, under the framework of Skill Acquisition Theory (Anderson et al., 2004; DeKeyser, 2015) and Skill Retention Theory (Kim et al., 2013).

In addition to the effects of temporal distribution of practice in L2 Mandarin word learning, this dissertation also attempts to explore the learning processes underlying L2 Mandarin word learning by NSs of nontonal languages under different practice distribution conditions. As learning to orally produce Mandarin disyllabic words is a complex task that may draw on not only declarative memory ability, but also procedural memory ability, working memory capacity, and musical aptitude, this

dissertation attempts to scrutinize the roles of individual differences in these aptitudes for the learning and retention of the learned knowledge and skills under different practice distribution conditions.

Chapter 2: Review of the Literature

2.1 Empirical studies on the distribution of practice

Research on temporal distribution of practice has a history of more than a century beginning with Ebbinghaus (1885/1964) and has since become one of the major research topics in learning and memory research. Not only of interest to cognitive psychologists, temporal distribution of practice is also of great interest to researchers in more applied fields, such as educational psychology, athletic training, surgical skills training, musicians' practice, and foreign language learning, due to its potential of practical utility.

In experimental psychology, *massed practice* refers to the conditions in which "repeated study opportunities occur in immediate succession", while *spaced* or *distributed practice* refers to the conditions in which "repetitions are spaced or separated by time and/or other events" (Toppino & Gerbier, 2014, p. 115). The term *spacing effect* refers to "enhanced learning during spaced as compared with massed study episodes for a given item" (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, p. 354). The term *lag effect* refers to "comparisons of different levels of spacing, either differing numbers of items or differing amounts of time" (Cepeda et al., 2006, p. 355). *Distributed practice effect* has been used as a generic term referring to "any finding in which a longer interval between successive study opportunities produces better performance on the final memory test than a shorter interval" (Toppino & Gerbier, 2014, p. 115), encompassing both spacing and lag effects, without distinguishing them (see Cepeda et al., 2006, p. 355).

In addition to being treated in an absolute sense as either massed or distributed (Cepeda et al., 2006), temporal distribution of practice has also been treated on a continuum from more massed to more distributed in studies with multiple spacing conditions. The term *massed practice* has also been used more liberally for conditions that are not necessarily massed in the absolute sense (with repetitions without any intervening events/time) but are massed in relative terms in that these conditions are comparatively more concentrated than the other condition(s) that is/are more spaced or distributed.

2.1.1 Cognitive and educational psychology

A vast body of literature in cognitive psychology has investigated whether a longer time interval between successive study opportunities produces better memory retention of the studied materials than a shorter interval (for review, see Cepeda et al., 2006; Toppino & Gerbier, 2014). There is substantial evidence that the *spacing effect*, i.e., the benefits of distributed practice (when repeated study opportunities are spaced by time and/or events) over massed practice (when repeated study opportunities occur in immediate succession), is robust and reliable (Cepeda et al., 2006; Toppino & Gerbier, 2014). The meta-analysis conducted by Cepeda, Pashler, Vul, Wixted, and Rohrer (2006), which is based on 839 assessments of distributed practice in 317 experiments located in 184 articles on verbal memory recall (of various materials, such as trivia facts, paired associates, paragraphs, faces, or objects), demonstrated convincingly that spaced presentations (with a time lag of 1s or longer) led to markedly better performance in verbal memory recall in the final test than massed presentations (with a time lag shorter than 1s). In addition, from the data they

gathered, they "failed to find any evidence that the [spacing] effect is modulated by retention interval" (Cepeda et al., 2006, p. 365). Note that the retention intervals varied from less than a minute to more than a month. In other words, the existing empirical evidence suggests that the spacing effect is robust regardless of the length of retention intervals in verbal memory tasks.

While the *spacing* effect (i.e., benefits of distributed over massed presentations) is robust and reliable, massed practice in the absolute sense, i.e., repetitions occurring in immediate succession without any intervening time or event, rarely happens in real life. Learning events in real life are almost always spaced or interleaved. Thus, it is of greater practical importance to investigate the *lag* effect by comparing the effects of different levels of spacing, with the goal of searching for the optimal spacing for learning and long-term retention. With regard to the lag effect, Cepeda et al.'s (2006) synthetic analyses of a large number of studies on verbal memory recall suggested that memory retention performance is affected by both inter-study interval and retention interval. Specifically, the analyses suggest that the effect of inter-study interval on final-test performance seems to be nonmonotonic: when increasing inter-study interval, retention performance improves; however, further increasing inter-study intervals results in reduction in retention accuracy. In addition, their analyses also suggested that the optimal inter-study interval that produces maximal retention increases as the retention interval increases.

Building on the findings of this meta-analysis by Cepeda et al. (2006), recent experimental studies began to seek the optimal spacing of practice for different levels of long-term retention that are educationally meaningful (Cepeda et al., 2009;

Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Rohrer & Pashler, 2007). The web-based experiment reported in Cepeda et al. (2008) is perhaps the most systematic examination of spacing effects for long-term retention. In order to reveal the interaction of ISI and RI, this study designed 26 experimental conditions formed by the different combinations of ISI (which varies from 0 to 105 days), and RIs (i.e., 1, 5, 10, and 50 weeks). In this experiment, 1,354 participants were taught a set of 32 obscure trivia facts in the first session until they reached a criterion of one perfect recall for each fact. After a prescribed ISI, participants came for a second session during which they were tested twice on each fact with feedback given. Then, after a prescribed RI, they were tested on each of the facts without feedback. The results of this study documented the existence of nonmonotonic lag effects, that is, at a given RI, an increase in ISI causes memory recall to first increase and then decrease. In addition, as RI increases, the optimal ISI increases as well. Finally, as RI increases, the ratio of optimal ISI to RI declines. The data of this study suggested that the optimal ratio of ISI to RI declined from about 20% to 40% for a 1-week test delay to about 5% to 10% for a 1-year test delay. Cepeda et al. (2009) reported two experiments that examined the effects of gap durations between two study sessions on subsequent test scores after a delay. Experiment 1 was conducted in a lab setting (with each participant trained and tested individually), in which participants learned 40 foreign words (Swahili-English word pairs) in two study sessions, with ISI varying among 5 min, 1 day, 2 days, 4 days, 7 days, and 14 days, and were tested after a test delay, with a RI of 10 days. Experiment 2 was conducted in a simulated classroom setting in groups of 1-6 participants. The task involves the learning of some trivia

(i.e., 23 not-well known facts in Part A and 23 not-well-known objects in Part B) in two sessions, with ISI varying among 5 min, 1 day, 7 days, 28 days, 84 days, and 168 days, and participants were tested after an RI of 168 days (i.e., 24 weeks). Both experiments demonstrated nonmonotonic lag effects; in addition, it was found that the optimal spacing was 1 day for the 10-days test delay (with the ISI/RI ratio of 10%) and was 28 days for the 168-days delay (with the ISI/RI ratio of 16.7%). Of particular interest to the present study, the results of both studies suggest that optimal spacing of practice for memory recall seems to be determined by the ratio of ISI and RI. Rohrer and Pashler (2007) suggested that the optimal ratio of ISI/RI for best retention ranges from 10% to 30%.

As the distributed effects have been well established in cognitive psychology, researchers have attempted to explore whether these distributed effects, well established in the laboratory settings using simple verbal memory tasks, can be generalized to real-life classroom settings and/or to other types of learning. A good number of classroom studies in the educational contexts have demonstrated the spacing and lag effects with regard to memory retention of verbal or factual materials, over a wide range of age groups (e.g., Bloom & Shuell, 1981, with high school students on the learning of foreign French vocabulary; Carpenter, Pashler, & Cepeda, 2009, with 8th graders on the learning of history facts; Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014, with 6th graders on the learning of foreign vocabulary; Seabrook, Brown, & Solity, 2005, with 1st grade children on the learning of basic literacy/reading skills; Sobel, Cepeda, & Kapler, 2011, with 5th graders on the learning of unfamiliar vocabulary). Sobel et al. (2011) compared the effect of

distributing two study sessions with an ISI of 1 week with that of massing them with an ISI of less than 1 minute on the retention of learned new words in 5th grade children in a classroom context after a RI of 5 weeks, and found that 1-week spacing produced superior retention than the massed condition. Küpper-Tetzel et al. (2014) examined lag effects in foreign vocabulary learning in a secondary school setting. Sixth graders learned and then relearned a set of 26 German-English word pairs with ISIs of 0, 1, or 10 days, and then tested 7 or 35 days after the last (second) learning session. They found that for the 7-day RI, the optimal ISI was 1 day, and shorter or longer ISIs led to lower performance (i.e., the nonmonotonic lag effect). For the 35-day RI, performance benefited from ISIs of both 1 day and 10 days. Küpper-Tetzel et al. concluded that the optimal ISI for the RI of 35 days was located "beyond a 1-day lag, with a 10-day lag leading to comparable benefits for memory performance." (p. 383), and suggested that the discrepancy between this finding and that of Cepeda et al. (2008) who found a significant increase in performance for ISIs from 0-day to up to 11-day for the 35-day RI, might be due to learner characteristics, such as working memory and forgetting rates, since Küpper-Tetzel et al. worked with children but Cepeda et al. with adults.

While the majority of the studies on distributed effects in either the laboratory or classroom setting have used verbal or factual materials as the to-be-learned stimuli, which requires no more than simple retrieval from memory, a number of studies have started to explore whether the distributed effects can be generalized to higher-level learning that goes beyond simple memory retrieval, such as inductive category learning (e.g., Kang & Pashler, 2012; Kornell & Bjork, 2008; Vlach, Sandhofer, &

Kornell, 2008; Wahlheim, Dunlosky, & Jacoby, 2011; Zulkiply & Burt, 2012; Zulkiply, McLean, Burt, & Bath, 2012), the learning of scientific concepts (e.g., Gluckman, Vlach, & Sandhofer, 2014; Kapler, Weston, & Wiseheart, 2015, on meteorology; Vlach & Sandhofer, 2012) and mathematics (e.g., Rohrer & Taylor, 2006, 2007, on permutation; Yazdani & Zebrowski, 2006, on plane geometry) (see Kapler et al., 2015, for a review). All the above-cited studies on inductive learning found that spaced presentation of stimuli leads to better discrimination of categories than massed presentation. For the learning of scientific concepts and mathematics, which involves mainly deductive learning, those studies also revealed that distributed practice led to better final test performance after a delay than more massed practice. In Rohrer and Taylor (2006), college students learned a mathematical procedure (i.e., permutation) either in a spaced condition (2 sessions separated by 1 week, 5 problems for each session) or in a massed condition (10 problems in a single session), and returned for a final test either after 1 week or after 4 weeks. It was found that (a) the massed group and the spaced group performed equivalently in the first session, (b) for the second session, the massed group averaged 94% in percent accuracy while the spaced group averaged only 85%, and (c) the two groups' performances were not significantly different at the 1-week test delay (ISI/RI ratio was 0% for the massed and 100% for the spaced); however, the spaced group (with an ISI/RI ratio of 25%) significantly outperformed the massed group (0%) at the 4-week test delay. In Kapler et al. (2015), undergraduate students attended a simulated university lecture where they learned some natural science concepts in meteorology. Participants reviewed the materials either 1 day or 8 days after the lecture and completed a final test five weeks

(35 days) after the review. It was found that for 5-week retention, the group that reviewed the materials 8 days after the lecture (ISI/RI: 23%) outperformed the other group who did the review only 1 day after the lecture (ISI/RI: 3%). Therefore, the optimal spacing ratio (i.e., ISI/RI between10% to 30%) that is generated from memory research (e.g., Rohrer & Pashler, 2007) seems to be able to generalize to higher-level learning. It is worth noting at this point that performance in higher-level learning is measured in accuracy (percentage correct) only, without taking into consideration of performance fluency or speed.

With respect to how to explain the well-documented distributed effects (including the spacing and lag effects), many theories have been proposed over the years, but not much agreement has been reached regarding the underlying mechanisms of these effects. Toppino and Gerbier (2014) present a recent review and evaluation of the major theoretical accounts of the distributed effects. According to them, almost all of the proposed theories rely on the following three basic mechanisms: deficient-processing mechanisms, encoding-variability mechanisms, and study-phase retrieval mechanisms.

Deficient-processing theories (Hintzman, 1976; Jacoby, 1978) posit that following the first presentation of an item, there is "a refractory period during which learners temporarily are unable or unwilling to process a second presentation to the extent that it is redundant with the first" (Toppino & Gerbier, 2014, p. 122). Therefore, the second repetition of a massed item does not receive effective processing, which explains the poorer memory of massed items. Toppino and Gerbier (2014) pointed out that deficient processing, which is attributed to short transitory

processes, may be able to explain the spacing effect; however, it does not seem to be able to explain the lag effect with longer degrees of spacing.

Encoding-variability theories (e.g., Genberg, 1979) emphasize the role of contextual variability during encoding in facilitating subsequent retrieval during testing. Variable encoding is assumed to facilitate later memory retrieval because of the idea that "the more different ways a stimulus or event has been encoded, the more different ways the target information can be found or access during retrieval" (Toppino & Gerbier, 2014, p. 123). Therefore, distributed practice results in better memory recall than massed practice because items presented with increasing lag are more likely to be encoded with different contextual information and therefore richer memory traces for later recall.

Study-phase retrieval theories (e.g., Thios & D'Agostino, 1976) argue that in the study/practice phase, during the second occurrence of an item, if its first occurrence can be successfully retrieved from memory, it strengthens and improves memory of the item. The benefits of successful retrieval during practice increases with lag (i.e., inter-study intervals), because with longer lags, successful retrieval is more effortful, which in return leads to better performance on the subsequent test. However, if the lags become too long and lead to failure in study-phase retrieval, the later repetition of the item then has little beneficial effect, and therefore results in poorer later test performance.

Due to the lack of consensus with regard to the mechanisms underlying distributed practice effect, and the fact that this phenomenon may be more complex than any one of the above basic mechanisms could explain, researchers have started

to take hybrid approaches and consider multiple mechanisms. One of such recent models, the multiscale context model (MCM) (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009) combines the search of associative memory (Raaijmakers, 2003), which encompasses the assumptions of both study-phase retrieval and contextual variability, and the predictive utility theory (Staddon, Chelaru, & Higa, 2002), which is based on the assumption that memory is limited in capacity and therefore to achieve optimal performance, memories should be erased if they are not likely to be needed in the future. Predictive utility refers to the idea that "the time that elapses before the re-encounter of information (i.e., the lag) determines for how long this information will be maintained in memory for the future" (Küpper-Tetzel & Erdfelder, 2012, p. 38). That is, "if the to-be-learned material is relearned after a long lag, our memory system will store and, importantly, maintain the material for a longer period of time. By contrast, if the lag is short, the material will be available for a short time only" (Küpper-Tetzel & Erdfelder, 2012, p. 38). The MCM therefore makes additional hypotheses regarding the maintenance processes of memory.

The mechanisms or models reviewed above seem to emphasize different memory processes (such as encoding, retrieval, and maintenance) that may be responsible for the distributed practice effects. A recent study by Küpper-Tetzel and Erdfelder (2012) investigated to what extent the nonmonotonic lag effects can be attributed to memory encoding, memory maintenance, or memory retrieval, based on their experimental data gathered from university students learning foreign vocabulary by using a multinomial processing tree model for free-then-cued-recall data. They

concluded that “the lag effect trends are mainly driven by encoding and maintenance processes rather than by retrieval mechanisms” (p. 37).

To summarize, in cognitive and educational psychology, the distributed practice effects (including the spacing and lag effects) have been consistently found across a wide range of learning tasks (verbal memory and higher-level learning) and in both laboratory and classroom settings. More research is absolutely required for a better understanding of the underlying mechanisms of the distributed effects. In addition, more empirical studies on distributed effects are still needed for higher-level learning, and real-life learning tasks that go beyond simple retrieval of factual information and additionally involve the manipulation of retrieved information and/or the application of the learned rules or sequences which typically characterize real-life skill learning that are more complex in nature. While the optimal spacing, i.e., the range of ISI/RI ratio (10% - 30%) seems to be compelling for memory retention of factual information, more empirical research is need to see whether the same optimal ISI/RI ratio can be extended to other types of learning, such as perceptual-motor skill learning, cognitive skill learning, etc. As real-life skills are typically complex and may involve learning of different types of knowledge or sub-skills, it remains a question how to best distribute practice time for optimal acquisition and retention of complex skills.

Moreover, as skill learning is not just about accuracy, i.e., whether procedures are correctly performed, skill fluency, i.e., how fluent and fast procedures can be performed correctly, also matters. While studies reviewed in this section only used accuracy in percent correct as the only outcome measure, probably due to the fact that

the studies on verbal memory solely focus on whether learned materials can be recalled in verbal or declarative memory and the studies on higher-level learning care only about whether the correct solutions can be reached, future studies on skill learning should also take speed of performance as an outcome measure.

2.1.2 Motor skill learning

Research on distributed practice effects on motor skills learning has a long history. In the initial meta-analysis of 47 articles (52 effect sizes) on the effects of distribution of practice on motor skill learning, Lee and Genovese (1988) defined distribution conditions in terms of the length of the inter-trial interval (ITI), with massed practice being the condition of the shortest ITI interval in a single practice session and distributed practice being the condition with the largest ITI interval. This meta-analysis focused solely on studies using psychomotor tasks, or simple motor tasks, such as rotary pursuit tracking, stylus mazes, inverted alphabet printing, mirror tracing, balancing, climbing, and basketball shooting (Lee & Genovese, 1988). The authors found that distributed practice enhanced not only acquisition (i.e., performance on the final trial for practice; $d = 0.96$) but also retention (i.e., performance after a retention interval has elapsed since the completion of the practice sessions; $d = 0.53$) compared to massed practice.

Lee and Genovese (1988) noted that almost all previous studies had examined the learning of *continuous* motor skill tasks (i.e., tasks with "prolonged time spent on the task", p. 284), with extremely few on the learning of discrete motor tasks (i.e., tasks that are "relatively rapid from initiation to completion", p. 284). In their search of literature to be included in their meta-analysis, they found only one study that

examined the distributed practice effects with a discrete motor task. This study (Carron, 1969) used a peg turn task, in which subjects were asked to "pick up a small dowel, turn it upside down, and reinsert the dowel into a small hole" as fast as possible (cited from Lee & Genovese, 1988, p. 284). The trial length was about 1300-1700ms; the inter-trial interval was 300ms for massed condition and 5s for the distributed condition. It was found that "massed condition resulted in moderately better learning, as measured on a retention test 2 days later" (cited from Lee & Genovese, 1988, p. 284).

Lee and Genovese (1989), using two versions of a movement timing task (the goal of the task was to learn to "move between two metal plates in as close to 500 ms as possible", p. 60), one version being *discrete* (which involved only one timing estimate per trial) and the other *continuous* (which involved 20 successive estimates per trial), investigated whether the effects of distribution of practice differ on the acquisition and retention of discrete versus continuous motor tasks. In line with the previous literature, on the continuous version of the task, the distributed practice condition (ITI = 25s) was found to produce better acquisition and retention (with two RIs: 10 min and 7 days) performance than the massed practice condition (ITI = 0.5s). When it comes to the discrete version of the task, on the contrary, the massed practice group outperformed the distributed practice group in both acquisition performance and retention performance.

Another study that found an advantage of massed practice over distributed practice on the retention of discrete motor skills is a recent study by Panchuk, Spittle, Johnston, and Spittle (2013), who also manipulated inter-trial interval to set apart

massed versus distributed practice conditions. In this study, Panchuk et al. investigated the effects of practice distribution on the acquisition and retention of a discrete sport skill, the Australian Football handball pass. All participants practiced the handball 50 times (5 blocks x 10 repetitions), either in a massed (ITI = 1 sec) or a distributed (ITI = 30 sec) condition with a 2-min rest between blocks. Performance was assessed in a pre-test, immediate retention test (RI = 10 min) and a delayed retention test (RI = 2 weeks). In terms of between-group comparisons, performance accuracy was not significantly different between the two groups on the pretest and delayed retention test; however, the distributed group performed significantly better on the immediate retention test. As for within-group comparisons, the massed group showed significant improvement in performance accuracy from pre-test to immediate and delayed retention tests, with no significant drop from immediate retention to delayed retention. Performance in the distributed group also showed a significant increase from pretest to immediate retention, but the scores dropped significantly from immediate retention to delayed retention, and there was no significant difference between scores of the pretest and the delayed retention. The authors concluded that "massed practice of a discrete sport skill may lead to a better retention of learning over a two-week period" (p. 751).

It is worth noting that the above reviewed studies including the meta-analysis all manipulated the practice distribution conditions in terms of relatively small *inter-trial intervals* (with shorter ITI being massed and longer ITI being distributed) in a single practice session; the overall pattern of the findings seems to be that the effect of distribution of practice in terms of ITI on motor skill learning may be modulated

by whether the motor task is continuous or discrete. For continuous motor tasks (i.e., with long trial lengths), distributed practice (in terms of longer ITI), has been proved to enhance both acquisition and retention of the learned motor skills. As for discrete motor tasks (i.e., usually with very short trial length, in milliseconds or seconds), there has been some evidence showing that massed practice (with an ITI of no more than 1 second) may lead to better retention of the learned skills than distributed practice (with an ITI of several seconds or up to 30 seconds), with a retention interval up to two weeks.

A few studies on motor skills learning have manipulated the spacing variable in terms of relatively longer *inter-session intervals*, with the goal of examining the comparative effects of spacing practice sessions across days versus within days (Dail & Christina, 2004; Shea, Lai, Black, & Park, 2000; Simmons, 2012). Shea et al. (2000) investigated the effects of distributing practice session across days relative to within days in two experiments, with Experiment 1 on a continuous balancing task, i.e., stabiliometer, and Experiment 2 on a discrete task, i.e., key-pressing timing. In both experiments, the practice sessions for one group were distributed within days and for the other across days (in Exp. 1: two practice sessions with an ISI of either 20 min or 24 hours; In Exp. 2: three practice sessions with an ISI of either 10 min or 24 hours). The RI was 24 hours after the completion of practice for both experiments. It was found in both experiments that the across-days acquisition groups started to outperform the within-day groups from the second practice session during practice, and also outperformed the within-day groups in the retention tests. Outside of the laboratory settings on simple psychomotor skills, Dail and Christina (2004) examined

the effect of practice distribution on a discrete sport skill, i.e., golf putting. 240 putts were either massed in one single session or distributed in four sessions, 60 putts per session, one session per day on four consecutive days. One third of the participants of each group returned for a retention test with an RI of 1 day, another one third with an RI of 7 days, and the final one third with an RI of 28 days. It was found that the group with practice distributed across days performed more proficiently than the massed group during both the remaining acquisition phase (since session 2) and the retention phase. A recent study, Simmons (2012), attempted to examine the roles of wake-based rest and sleep-based rest on the learning of a complex motor skill, i.e. piano keyboard playing. Three groups of non-pianist musicians (about 10 for each) learned a 9-note sequence in three practice sessions, with ISIs of 5 min (massed), 6 hours (distributed; within-day), and 24 hours (distributed; across-days), respectively. No retention test was administered. The results revealed that wake-based rest only improved performance speed but not performance accuracy, whereas sleep-based rest (ISI = 24 hr) improved not only speed but also accuracy. The authors of these three studies all attributed the enhanced performance in distributed practice across days (ISI = 24 hr) over massed practice within days to memory consolidation, i.e., the neurophysiological processes that transformed memories from relatively unstable states into relatively permanent form that are resistant to interference and forgetting (Shea et al., 2000; Simmons, 2012).

The effects of practice distribution have also been investigated in the learning or training of surgical skills, a type of complex fine-motor skills that require a heavy cognitive preparation (Mackay, Morgan, Datta, Chang, & Darzi, 2002; Moulton,

MacRae, Graham, Grober, & Reznick, 2006). Mackay et al. (2002) reported that in the learning of a laparoscopic surgical skill, novice subjects trained in a distributed practice condition (20 min training time separated in four 5-min blocks with 2.5-min intervals in between) performed significantly better than those trained in the massed practice condition (20 min in a single session) on the retention test (RI = 5 min). Moulton et al. (2006) reported that in the learning of a microsurgical skill (i.e., microvascular anastomosis), surgical residents trained in the distributed practice condition (i.e., 4 sessions, one per week) significantly outperformed those trained in the massed practice condition (i.e., 4 sessions on 1 day) on the retention test (RI = 1 month). The findings of both studies lend support to distributed practice rather than massed practice in surgical skill training.

To summarize, when *inter-trial interval* was used as the criterion to distinguish massed versus distributed practice, typically within a single practice session, the effects seem to be modulated by whether the motor task is continuous or discrete. For continuous motor skills, there has been substantial evidence that distributed practice enhances both acquisition and retention of the learned skills. However, for discrete motor tasks, there has been some consistent evidence that massed practice leads to better retention of the learned skills than distributed practice. When *inter-session interval* was manipulated to distinguish massed versus distributed practice, the existing studies have only compared distributed practice across days (i.e., ISI $\geqslant$1 day) with massed practice within days (i.e., ISI < 1 day), showing an advantage for the former over the latter. However, none of the studies I have gathered compared ISIs larger than 1 day with an ISI of 1 day. Empirical studies on motor skill

learning have not systematically manipulated a wide range of ISIs and RIs. More such empirical studies are absolutely required before a complete picture can be revealed about how to structure practice time for optimal acquisition and retention of different types of motor skills.

2.1.3 Second Language Acquisition Research

In SLA, the effects of temporal distribution of learning, practice, or instruction hours in L2 learning have received some attention. One line of research is second/foreign language program evaluation studies that attempted to compare the relative effectiveness of intensive/compact courses with regular distributed foreign language courses (e.g., Collins, Halter, Lightbown, & Spada, 1999; Collins & White, 2011; Lapkin, Hart, & Harley, 1998; Lightbown & Spada, 1994; Serrano, 2011; Serrano & Muñoz, 2007).

Admittedly, the early studies (Collins et al., 1999; Lapkin et al., 1998; Lightbown & Spada, 1994) suffer from some methodological problems in that the intensive versus distributed courses differed in other ways than temporal distribution, such as instruction hours, exposure time outside of school, or learner ability. These studies all found an advantage for intensive courses compared to regular distributed courses, even for long-term retention. However, these results are hard to interpret because the other factors also systematically favored the intensive group (see also Bird, 2010; Rohrer, 2015). For instance, in Lightbown and Spada (1994), participants in the intensive group reported more outside exposure to the target language than the distributed group before the final tests. In Collins et al. (1999), the intensive program provided more instruction hours than the regular distributed program; in addition, the

intensive program was “limited to students who had above-average academic ability” while the regular program included students with “a wide range of ability” (p. 660). These confounding variables make the results of these studies hard to interpret regarding the effects of temporal distribution.

Recent studies that carefully controlled the total amount of study time and manipulated the distribution of instruction hours (Collins & White, 2011; Serrano, 2011; Serrano & Muñoz, 2007), however, still found no advantage for distributed schedules over massed schedules. On the contrary, they seem to demonstrate an overall advantage for massed practice in L2 learning. Using a pretest-posttest design (with pretest at the beginning and posttest at the end of the course), Serrano and Munoz (2007) found that EFL Spanish-speaking college students in the intensive and semi-intensive programs (intensive: 5 hours/day, 5 days/week, over 5 weeks; semi-intensive: 10 h/week, 2.5 h per day, M to Th, over 11 weeks; or 8 h/week, 2 h per day M to Th, over 15 weeks) made significant improvement in proficiency tests in listening, reading and grammar (as measured in sentence conversion), but students in the extensive program (4 hours/week in two days over 7 months) did not. The three programs had the same amount of instruction hours, i.e., a total of 110 hours. Using the same design, while adding a dimension of proficiency levels and tests in writing and speaking skills, Serrano (2011) found that intermediate-level students made significantly more gains in proficiency tests in intensive programs than in extensive programs, and comparable gains in both groups in fluency, complexity, and accuracy measures in written and oral production tasks. For advanced-level students, the intensive and extensive groups made comparable gains in proficiency tests as well as

written and oral production tasks. Working with ESL French-speaking children at Grade 6 (aged 11-12 years), Collins and White (2011) compared the relative effectiveness of two intensive programs, one with 400 instruction hours concentrated into a 5-month block and the other with the same amount of hours distributed in a series of intensive exposures over a 10-month academic year. The language development was compared between the two groups four times (at 100-hour intervals during their respective programs) via a battery of aural and written comprehension and oral and written production measures. They found that "there were significant differences between the two groups on 7 of the 20 between-group comparisons… Of the significant differences, six showed advantages for the concentrated group, compared to just one for the distributed group" (p. 125). Collins and White concluded that there was "no clear learning advantage for either concentrating or distributing the intensive experience." (p. 106).

To recap, the findings of these non-confounded studies (Collins & White, 2011; Serrano, 2011; Serrano & Muñoz, 2007) seem to suggest that intensive programs are more effective than distributed or extensive programs on L2 acquisition of certain aspects of a language, at least at some stages. None of these studies demonstrated evidence for any advantage for distributed over intensive programs, which seems to contradict the findings from the cognitive psychology literature. Three possible explanations for the discrepancy of these findings from the two research paradigms have been suggested in the literature. First, as pointed out by Serrano (2011), in L2 classes, repetitions of vocabulary and grammatical structures are always spaced, with intervening items or materials. Therefore, "language learning

in both the intensive and the regular programs can be said to be distributed in cognitive psychology terms" (Serrano, 2011, p. 123). Following the study-phase retrieval theories, repeated study opportunities of the same item, concept or structure in the extensive programs may be too spaced out that it may lead to failure in retrieving the item or the rule for the target structure. This failure of retrieval makes proceduralization of L2 skills hard (DeKeyser, 2007b). In the intensive programs, repetitions of the same item or structure are still spaced; however, the spacing is not too wide. Therefore, the earlier learned knowledge, either vocabulary or grammatical rules, is more readily available for retrieval and therefore for proceduralization (Serrano, 2011). Thus, this discrepancy in the meaning of massed versus distributed in cognitive psychology and intensive versus extensive schedules in language programs may be able to account for the seemingly contradictory findings from the two different research paradigms. Second, the types of target skills examined in these FL program evaluation studies and those in cognitive psychology are different, as hinted by Collins and White (2011). While studies in cognitive psychology focus mostly on simple memory recall of verbal or factual information, the L2 program evaluation studies primarily focus on global L2 skills, such as listening, speaking, reading and writing, that are more complex by nature, involving the processing and integrating of multiple types of knowledge (lexical, grammatical and phonological). It has been suggested in Donovan and Radosevich's (1999) meta-analysis that increased task complexity may attenuate distributed practice effects. Lastly, none of the nonconfounded studies favoring intensive L2 learning included a delayed posttest, as pointed out by Rohrer (2015). In other words, these studies assessed the effect of

temporal distribution (intensive vs. distributed) on learning by the end of the courses; however, it remains unknown whether the learning effects are durable in the long term. On the other hand, retention is the focus of the studies in cognitive psychology. These are three explanations that can possible account for the seemingly contradictory findings from these two research paradigms.

Another line of research in SLA has examined the effects of temporal distribution on L2 learning of discrete items, such as vocabulary (e.g., Bloom & Shuell, 1981; Cepeda et al., 2009, Experiment 1; Karpicke & Bauernschmidt, 2011; Küpper-Tetzel & Erdfelder, 2012; Küpper-Tetzel et al., 2014; Nakata, 2015; Pavlik & Anderson, 2005; Schuetze, 2015) and grammatical structures (Bird, 2010; Miles, 2014; Rogers, 2015; Suzuki, 2017a, 2017b, Suzuki & DeKeyser, 2015, 2016). For L2 vocabulary learning, previous experimental studies have mainly focused on the memorization of paired associates, and the results all demonstrated the distributed practice effects consistent with the findings from cognitive psychology. Participants in these studies were typically visually presented L2-L1 word pairs during the study phase, and when it comes to testing, they were typically presented with the L2 words visually, and were asked to recall the meanings of the words by typing the L1 translations (e.g., Cepeda et al., 2009, Experiment 1). It is worth noting that the word learning tasks in these studies involve only declarative memory recall, i.e., simple verbal memory retrieval, without the involvement of auditory word perception or oral production, which are in fact vital aspects of L2 word learning, if the learning purpose is for real L2 communication.

L2 grammar learning is a type of higher-level learning that goes beyond simple verbal memory recall and additionally involves manipulation of retrieved information and/or application of the learned rules in new contexts, especially in instructed foreign language classroom settings. L2 grammar learning is therefore more abstract and complex. Only a few recent studies have attempted to explore whether the distributed practice effects that are well established in L2 vocabulary learning and cognitive psychology can be generalized to L2 grammar learning. All the existing studies have been conducted in foreign language contexts with university students, using a pretest-posttest design.

Bird (2010) compared the effects of distributed versus massed practice schedules on L2 learning of English tense and aspects (i.e., simple past tense, present perfect, and past perfect) by NSs of Malay. The inter-session interval (ISI) was 3 days for the massed group and 14 days for the distributed group. Both groups went through 5 study sessions at designated schedules and were tested at a short (7-day) RI and a long (60-day) RI. This study was conducted in a classroom setting. For both practice and tests, students were given worksheets and told to judge whether the verb forms of the sentences were correct or not, if incorrect, to correct them. Results showed that both groups made significant and equivalent improvements on the short-term (7-day) delayed posttest. More importantly, however, the gains were only retained by the distributed practice group, but not by the massed group, on the long-term (60-day) delayed posttest, suggesting that the distributed learning condition yielded better long-term retention of the learned grammatical knowledge.

Miles (2014) examined the temporal distribution effects on L2 learning of English adverb placement by Korean university students. This study was also conducted in a classroom setting. A total of 65 instruction hours were massed into a single session for the massed learning condition, and distributed into three study sessions (40 min for the first, 10 min for the second, and 15 min for the third) over a 5-week period (1-week interval between the first and second study sessions and 4 weeks between the second and third study sessions) for the distributed learning conditions. The delayed posttests were conducted after a 5-week RI. Results showed that for error correction, starting at similar levels in the pretest, the two groups performed similarly on the immediate posttest, suggesting similar levels of performance at the end of study session(s) and similar level of gains through the study session(s) as measured by the test; on the delayed posttest, however, the distributed group outperformed the massed group, suggesting that the distributed learning condition yielded better long-term retention of the learned knowledge. For translation performance (written, from L1 to L2), there were no significant differences between the two groups either on the immediate posttest or on the delayed posttest.

Rogers (2015) examined the temporal distribution issue on L2 learning English cleft sentences by EFL learners in Middle East. The training materials were a total of 100 stimulus sentences (grammatical) in a comprehension check task. In a classroom setting, the distributed group went through the five study sessions (15 minutes each) with an ISI of 7 days, the massed group with an ISI of 2.25 days, and both groups completed a delayed posttest with an RI of 42 days (i.e., 6 weeks). The ISI/RI was 17% for the distributed condition and 5% for the massed group. Using the

accuracy scores from untimed written grammaticality judgment tasks (GJT) as the measure of learning outcome, the study found no significant differences between the two groups in the pretest and immediate posttest; however, the distributed group significantly outperformed the massed group in the delayed posttest after a 6-week delay. This study, therefore, again contributed evidence that distributed practice leads to better long-term retention of the learned grammatical knowledge.

Suzuki and DeKeyser (2015) examined the effects of temporal distribution on L2 proceduralization of a Japanese morphological structure, present progressive *–te,* by English NSs in a laboratory setting. The ISI was 1 day for the massed group and 7 days for the distributed group; both groups went through 2 training sessions and were tested after a short (7-day) RI and a long (28-day) RI. The ISI/RI ratio therefore fell into the optimal range (i.e., 10%-30%) identified in cognitive psychology for the 1-day ISI condition when tested at the RI of 7 days (14%), and for the 7-day ISI condition when tested at the RI of 28 days (25%). When the 1-day ISI group was tested at the 28-day RI (3%) or the 7-day ISI group was tested at 7-day RI, the ISI/RI ratios were suboptimal. Each training session consisted of four tasks – vocabulary learning, explicit grammatical explanation, auditory comprehension practice (auditory), and oral production practice. The learning outcomes were measured in two tests, a rule application test and a picture sentence completion test. The former test was used to assess learners' ability in using the morphological rules in converting the base forms of novel verbs into present progressive; the latter test was used to assess learners' ability in using the correct form of the practiced verbs to describe the actions of the seen pictures, which is more contextualized, meaning-oriented, and complex.

The learning outcome was measured in both accuracy (in percentage correct) and response time, operationalized as the duration of time from the presentation of the visual stimuli (either word or picture) to the end of the utterance. Results showed that, for accuracy measures, the two treatments, i.e., distributed versus massed practice, seem to result in comparable performance in both rule application and picture sentence completion tests across the board at the immediate posttest and the two delayed posttests. The speed measures, however, showed an advantage of the massed practice condition in the picture sentence completion task in the long-term retention test (28-day delay). In a follow-up exploratory analysis on the role of cognitive aptitudes in determining the effects of different practice distributions, Suzuki and DeKeyser (2016) found that language-analytic ability was only related to performance after distributed practice, whereas WM capacity was only related to performance after massed practice, in sentence completion.

A most recent study, Suzuki (2017a), attempted a conceptual replication and extension of Suzuki and DeKeyser (2015). The replication study, Suzuki (2017a), improved three aspects of the design of Suzuki and DeKeyser (2015): first, the replication study used a novel miniature language in order to control learners' prior knowledge about the target structure; second, the replication study increased the sample size to 60 participants (30 in each group); third, the replication study doubled the number of training sessions (i.e., 4 training sessions). The replication study also extended the previous study by adding a dimension of linguistic complexity of the target learning task. The ISI was 3.3 day for the massed practice condition and 7 day for the distributed practice condition, and two delayed post-tests were administered

for both groups, the first after a 1-week delay and the second with a 4-week delay. The replication study found a robust advantage for massed practice over distributed practice, i.e., the 3.3-day ISI group started to provide more accurate performance than the 7-day ISI group from the beginning of the third training session, and this advantage was maintained on both the 1-week and 4-week delayed posttests. As for utterance speed, there was no significant difference between the two groups in RT across the training phase or at the two delayed posttests. Linguistic complexity, as operationalized in that study, was not found to exert an influence on the effectiveness of the different practice distribution conditions. From a follow-up analysis of the ATI effects, Suzuki (2017b) found that procedural learning ability measured by the Tower of London task was only significantly associated with RT in the 3.3-day ISI group, but not in the 7-day ISI group.

The findings of Suzuki and DeKeyser (2015) and Suzuki (2017a) run counter to the findings of the other studies on L2 grammar learning (Bird, 2010; Miles, 2014; Rogers, 2015). The contradictory findings can be attributed to a number of factors. First, the much higher complexity of the learning tasks and the outcome tests might have attenuated the distributed practice effects. While the other three studies all used paper-and-pencil error identification and/or correction tasks to measure whether learners were able to retrieve the learned rules and apply them in appropriate contexts given presumably unlimited time, the learning tasks and outcome tests in Suzuki and DeKeyser (2015) and Suzuki (2017a) were more cognitively demanding because they required the execution of more cognitive processes (conceptualizing the speech, retrieving lexical items, applying a morphological rule, and finally articulation) in a

timed manner. Second, differences in skill acquisition stages may have contributed to the inconsistent findings. The other studies only required participants to detect and/or correct grammatical errors, which seem to focus mainly on the declarative stage of learning, perhaps with some incipient proceduralization. Suzuki and DeKeyser (2015) and Suzuki (2017) focused on the latter stages, i.e., proceduralization and incipient automatization. While distributed practice may work better for the retention of declarative knowledge, massed practice might be better for the acquisition of procedural knowledge. The other factors that may worth further exploring include the number of practice sessions, the ISI/RI ratio, the type of linguistic knowledge (e.g., receptive vs. productive) and linguistic knowledge domains (Suzuki & DeKeyser, 2015; Suzuki, 2017a).

In summary, research findings regarding the effects of temporal distribution of practice on L2 learning and retention have been inconsistent. The macro-level program evaluation studies have generally found an advantage for intensive programs over regular extensive programs; however, the outcomes for both groups were only compared at the end of the courses, and none of the well-controlled non-confounded studies have assessed long-term retention. For the laboratory experimental studies that have examined the learning of discrete L2 items, research findings from L2 vocabulary learning have been consistently favoring distributed over massed practice for long-term retention. For L2 grammar learning, three studies that focused on the learning of L2 grammatical rules and the application of the rules in offline grammaticality judgment tests found an advantage for distributed practice for long-term retention; however, Suzuki and DeKeyser (2015) and Suzuki (2017a) who

focused on L2 grammar in oral production found no such advantage for distributed practice, and on the other hand demonstrated an advantage for massed practice. A distinction between acquisition performance (i.e., performance at the end of the practice sessions) and retention performance (i.e., performance after a retention interval) seems necessary when interpreting the results. In addition, the nature and complexity of learning tasks and materials (simple memory retrieval of factual information, manipulation of retrieved information in offline tasks, versus online performance of complex skills, which involves the integration of multiple sources and/or types of knowledge) and stages of learning for complex skills (declarative, proceduralization, automatization) seem to be promising candidates to explain the inconsistent findings.

Finally, there was a methodological issue in the most recent L2 studies of distributed learning on grammar learning, including Bird (2010), Suzuki and DeKeyser (2015), and Suzuki (2017a). In these studies, RI was treated as a within-subjects variable, rather than a between-subjects variable. In other words, multiple delayed posttests (usually two, the first at a shorter RI, and the second at a longer RI) were administered to assess the retention of knowledge or skill within participants. While this within-subjects design for RI reduces the cost for data collection by cutting off the training sessions for half of the sample size, this design also introduces a confound. That is, testing at the short-term RI could have exerted an influence on retention performance at the longer-term RI, resulting assessment at the longer RI invalid/inaccurate. We are unknown or unsure at best about the extent of the effects of testing at the short-term delay on retention performance at the long-term delay, or

whether this effect differs depending on the practice schedule. Therefore, the results from these studies with this within-subjects design for RI are only valid/accurate for the effects of ISI and the short RI on retention performance; retention performance at the longer RI in these studies is contaminated by testing at the short RI. This confounding factor has been acknowledged in Suzuki (2017a). Only a between-subjects design for RI can solve this issue. This dissertation uses RI as a between-subjects variable, so that both the effects of ISI and RI (at both levels) can be accurately assessed.

2.1.4 Summary

While the distributed practice effects on verbal memory tasks, including foreign vocabulary learning, have been robust and consistent, the current research findings on the effects of temporal distribution of practice on skill learning, including L2 skills, are much clouded. The effects of distributed practice in simple memory recall tasks (i.e., on declarative memory retention) seem to be determined mainly by temporal variables, i.e., the ISI/RI radio. The distributed practice effects in skill acquisition and retention, however, seem to be mediated by a number of factors, including the complexity of the learning task (Donovan & Radosevich, 1999; Suzuki & DeKeyser, 2015), the stages of skill learning (declarative, proceduralization, and automatization) (Kim et al., 2013), whether the focus is on acquisition performance or retention performance (Donovan & Radosevich, 1999; Lee & Genovese, 1988), and the frequency of practice sessions (Suzuki & DeKeyser, 2015), in addition to temporal variables (ISI, RI and/or ISI/RI ratio) (Cepeda et al., 2006; Rohrer & Pashler, 2007). The nature of the type of knowledge to be learned (declarative,

procedural, versus perpetual-motor) has also been suggested recently by Paik and Ritter (2016) as a mediating variable of distributed practice effect (which will be introduced in the next section). Another factor, not explicitly pointed out in the literature, is the type of outcome measures, accuracy or speed, in light of Suzuki and DeKeyser's (2015) findings.

More empirical research is absolutely needed on the learning and retention of L2 skills in SLA, such as L2 grammatical skills in oral production, or L2 word skills that involves auditory perception and oral production in addition to memorization of word meanings, with learning outcome measured not only in accuracy but also in automaticity/speed. Such research can help seek optimal spacing for L2 skill acquisition and retention. Up till this point, no studies have examined temporal distribution issue on word learning that takes auditory perception, oral production, and L2 phonology into account. The present dissertation attempts to fill this gap by examining the temporal issue in a study of L2 learning of oral Mandarin word production.

2.2 Theoretical framework

2.2.1 Skill acquisition theory

Skill acquisition theory sets out to account for "how people progress in learning a variety of skills, from initial learning to advanced proficiency" (DeKeyser, 2007a, p. 97). Skill acquisition theory is therefore pertinent to second language acquisition if we agree that the ultimate goal for most L2 learners is to achieve advanced proficiency in a set of skills that enable them to comprehend and produce

the second language fast and efficiently for communication (DeKeyser & Criado, 2012). In the most widely accepted model of skill acquisition, i.e., the ACT model of the human cognitive architecture, or ACT-R (Adaptive Control of Thought – Rational) in its later versions (Anderson, 1993; Anderson et al., 2004; Anderson & Lebiere, 1998), Anderson claims that the learning of a variety of skills goes through a similar trajectory of development from initial representation of knowledge to highly skilled behaviors. That is, skills are typically initially learned as declarative knowledge (or "knowledge THAT", such as instructions, examples, or facts about general properties of objects). This initial learning is followed by a rapid stage of proceduralization (knowledge compilation), which leads to qualitatively different procedural knowledge (or "knowledge HOW", which encodes behavior) through initial practice. Note that proceduralization does not constitute any transformation of knowledge from one type to the other, but rather that "declarative knowledge, via practice, plays a causal role in the development of procedural knowledge" (DeKeyser, 2015, p. 103). Procedural knowledge is then fine-tuned and automatized over a long period of time through a large amount of practice (Anderson, 1993; DeKeyser, 2007a, 2015; DeKeyser & Criado, 2012). Learners' performance in the later stages is featured by a gradual decrease in both reaction time (RT) and error rate. This decrease as a consequence of practice is generally held to take the form of the power law (Anderson, 1993; Newell & Rosenbloom, 1981). This phenomenon, which has been repeatedly observed in the learning of many different skills, is referred to as the power law of practice in skill acquisition.

These three stages of skill acquisition that Anderson (1982) termed as "declarative", "knowledge compilation", and "procedural" in the ACT-R model roughly correspond to Fitts' (1964) "cognitive", "associative", and "autonomous" stages from a cognitive information-processing perspective. Despite variation in terminology, the cognitive processes underlying each of the three skill-acquisition stages are generally held to be the same across different research approaches. Anderson claimed that "skill acquisition starts out with a large cognitive component. With practice, that cognitive component decreases… with continued practice, the thinking component continues to decrease. Eventually, all cognitive involvement is squeezed out, and there is only an automated motor routine." (Anderson, 2000, p. 306-307).

Further along this line of theorizing, Ackerman (1988; Ackerman & Cianciolo, 2000) carried out a series of experiments on cognitive-motor skill acquisition designed to characterize the cognitive processes underlying each of the three stages of skill acquisition. Based on their empirical findings and previous research findings, Ackerman proposed an integrative theory that links the three stages of skill acquisition with cognitive ability determinants of individual differences in performance during skill acquisition. Ackerman specified that performance levels in the initial cognitive (or declarative) stage are associated with general intelligence, performance levels in the associative (or knowledge compilation) stage are related to perceptual speed ability, and performance in the autonomous (or procedural) stage are determined by psychomotor ability. Perceptual speed refers to "speed of consistent encoding and comparing symbols". This ability, i.e., "the facility and speed of

compilation of production systems" are considered to determine performance efficiency during the associative stage of skill acquisition (Ackerman, 1988, p. 290). Psychomotor ability refers to "the speed and accuracy of motor responding that are characteristic of psychophysical limitations of the human subject" (Ackerman, 1988, p. 291); this ability was found the only factor that determines performance at the autonomous stage during motor skill learning.

Using the same approach, i.e., by studying the correlations between cognitive ability measures and performance levels during skill acquisition, Beaunieux and colleagues (Beaunieux et al., 2006) attempted to characterize the three stages of cognitive procedural learning and identify their boundaries. In addition to training in a Tower of Toronto task, participants were also administered a battery of cognitive tasks designed to measure six cognitive abilities, i.e., general intellectual functions, working memory, episodic memory, executive functions, perceptual processing, and psychomotor abilities. The results confirmed the contribution of general intelligence for the cognitive stage, and the contribution of psycho-motor abilities during the autonomous stage, providing bases for locating the boundaries of the three stages of cognitive procedural learning. In addition, both working memory and episodic memory contributed to performance at the cognitive stage. Perceptual processing abilities, hypothesized to be related to performance at the associative stage, was not found to be specific to any learning stages, as its correlations with task performance remained significant and stable throughout the learning process.

In order to further characterize the distinct processes underlying the three stages of cognitive procedural learning, Hubert, Beaunieux and colleagues (Hubert et

al., 2007) conducted a positron emission tomography (PET) activation study using the Tower of Toronto task. They found the involvement of the prefrontal cortex, cerebellum, and parietal regions during the cognitive stage, which was interpreted as suggesting the use of problem-solving strategies during that stage. They found the involvement of the occipital regions during the associative and autonomous stages, which was interpreted as suggesting evidence for intervention of mental imagery. Finally, the activation of the anterior part of the cerebellum was found during the autonomous stage, and this was interpreted as providing support for the hypothesis that performance during the autonomous stage is determined by psychomotor abilities.

To sum up, the three stages of development during skill acquisition have been well theorized in skill acquisition theories from different perspectives. There has also been some empirical data from both behavioral studies and brain imaging studies supporting the segmentation of the three stages because distinct cognitive processes have been found to be involved during these different stages.

2.2.2 Skill retention theory

While it is important to explore best strategies for most efficient skill acquisition, it is equally important to explore best strategies for long-term skill retention because skill decay can happen with the passage of time. Based on current understanding of learning and retention in human cognition, Kim, Ritter and Koubek (2013) introduced a unified theory of skill retention that integrates learning with forgetting for a broader view of improved skill acquisition and retention that offers

practical implications as to how to best distribute practice at the different stages of skill acquisition for improved skill retention.

Kim et al.'s (2013) skill retention theory is based on Anderson's (1982) ACT-R model of skill acquisition; therefore, this skill retention theory also distinguishes the two types of knowledge (declarative vs. procedural), or knowledge types learned through declarative versus procedural memory, and the three distinct stages of learning, i.e., declarative, transitional (knowledge compilation), and procedural. Based on the premises that the forgetting mechanisms differ drastically for declarative and procedural memory (in that knowledge stored in declarative memory decays with the passage of time, while knowledge in procedural memory, i.e., production rules, does not decay with time, as explicitly specified in ACT-R), and that skill performance during the three stages of learning draws on different knowledge types (i.e., declarative, a mix of declarative and procedural, procedural), Kim et al. (2013) predicted that the degree and speed of forgetting will differ depending on the stage of skill acquisition that the learning stops at, and proposed that the optimal spacing of practice for skill acquisition is determined by learners' progression through the three stages of learning.

Figure 1 is the graph that Kim et al. (2013) used to illustrate their theory of skill retention. This figure depicts learning and forgetting curves across the three stages of learning. The continuous solid line represents the learning curve with continuous practice. The dashed lines represent the forgetting curves at the each of the three stages due to periods of inactivity, and the short solid lines represent the learning curves during later training. As can be seen from the slopes and heights of

the forgetting curves at the three different stages, the rate and extent of delaying differs across stages, with skill performance decay the fastest and the most when learning stops at the first stage, and skill performance decay the slowest and the least when learning stops at the third stage.

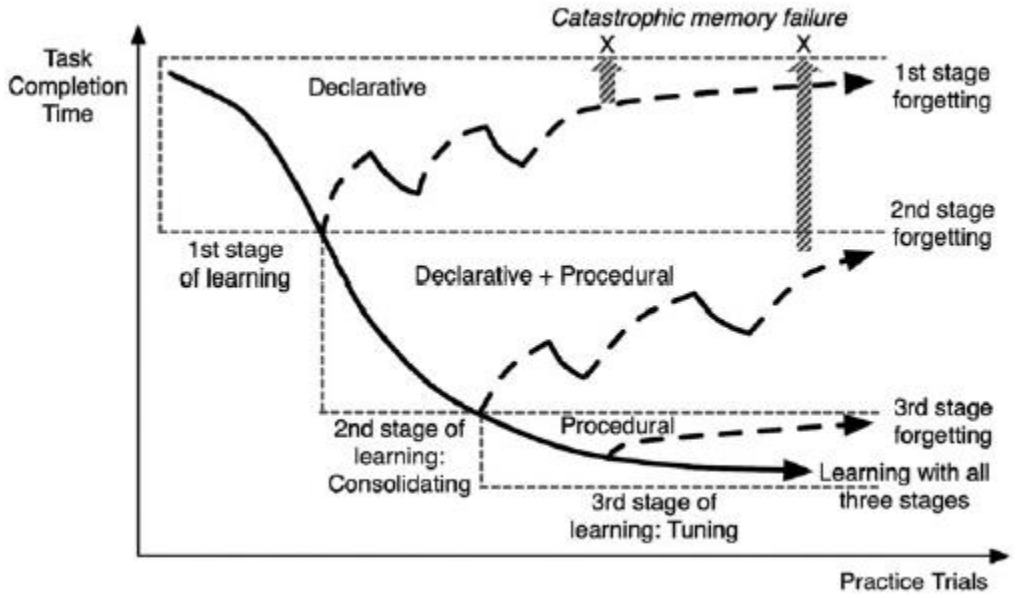


Figure 1. A graph depicting skill retention theory, showing learning and forgetting curves, from Kim, Ritter, and Koubek (2013), p. 26.

In the initial stage, i.e., the declarative stage, performance relies on declarative knowledge (knowledge in declarative memory). As declarative knowledge decays with lack of practice, a long period of inactivity may result in "catastrophic memory failure", that is, "a state where declarative memory items needed to perform the task cannot be retrieved from memory due to lack of practice (decay)" (p. 26). If the learning stops at this declarative stage, it is suggested that learners' performance may be optimized by distributed practice, of course with the inter-session intervals not as long as to result in catastrophic memory failure.

In the second learning stage (proceduralization), task performance relies on a mix of declarative and procedural knowledge; that is, performance draws on both declarative and procedural knowledge. At this stage, some declarative elements have been proceduralized with some practice, while some others have not yet been (fully) proceduralized. Progression in this stage moves from drawing on a mix with a larger portion of declarative knowledge to a mix with a larger portion of procedural knowledge. From a cross-sectional perspective, as the subskills could vary in their knowledge mix (take L2 oral production of Mandarin words from pictures for example, the subskill of remembering the picture-spelling mappings draws primarily on declarative memory, while the subskill of oral production from spelling is largely procedural), the slope of the forgetting curves could vary for the different subskills. As procedural memory is immune to decay while declarative memory decays with lack of use, if learning stops at this mixed stage, frequent practice of declarative knowledge, particularly of those elements that have not been proceduralized, is necessary to avoid catastrophic memory failures and to keep declarative knowledge active to support proceduralization because "declarative memories have to be active enough for new procedural rules to be generated" (p. 26). In addition, it is hypothesized that "if the learner's knowledge is about to move into the third stage (procedural knowledge), the learner's performance would be optimized by massed practice, which makes the declarative knowledge strong enough to proceduralise" (p. 30).

When it comes to the third stage of learning, i.e., the procedural stage, it is believed that "task knowledge is available in both declarative and procedural forms,

but procedural knowledge predominantly drives performance" (p. 27). Therefore, catastrophic memory failure is not likely to happen at this stage because declarative knowledge has been maximally proceduralized at this stage and procedural knowledge is robust. Skill retention theory further hypothesizes that with lack of practice, declarative knowledge may decay, but learners can still perform the task if "all the knowledge is proceduralized or is available in the environment and thus not forgotten with time" and "performing the task does not require new declarative inputs" (p. 27). Kim et al. (2013) note that although *procedural knowledge* does not decay in the ACT-R model, *skill* can decay, because "procedural knowledge, in a sense, can be and in some senses has to be primed by declarative knowledge" (p. 29). Therefore, forgetting of procedural knowledge can still happen if the declarative elements that are necessary for triggering procedural knowledge become inaccessible due to long periods of disuse. It is suggested that, while knowledge at this stage is mostly procedural, distributed practice would work the best to retain the declarative components and therefore the skill performance, while massed practice will not help.

Kim et al. (2013) further note that subcomponents of a skill may be learned and forgotten at different rates (see Figure 2 for an illustration); therefore "concerted and structured practice" is required to proceduralize each of the subskills (p. 27).

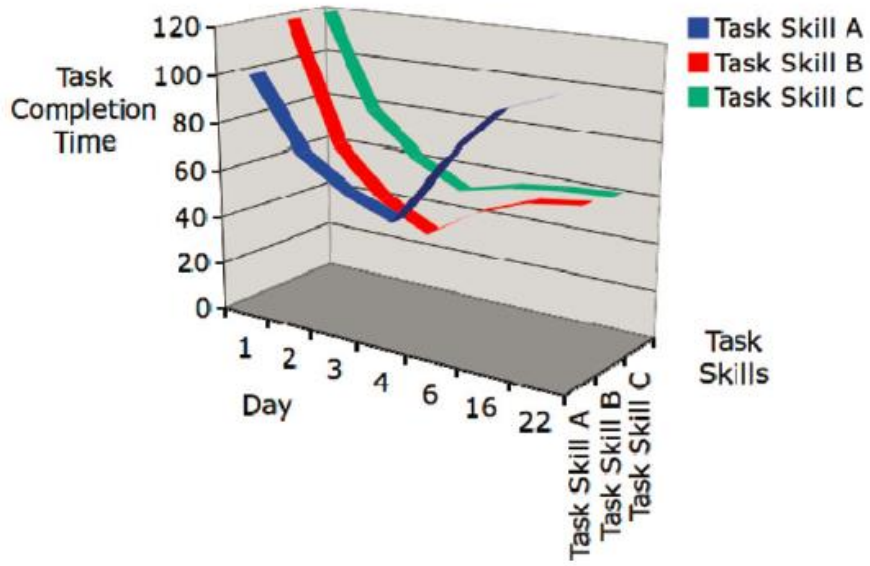


Figure 2. A graph showing the learning and forgetting of different subskills, from Kim, Ritter, and Koubek (2013), p. 27.

Skill retention theory provides an elaborated and more nuanced account for skill acquisition, forgetting, and retention. This theory suggests that the optimal spacing of practice should differ across the three different stages of learning because the task knowledge at the three learning stages is comprised of knowledge types with different structures (declarative, declarative + procedural, procedural), and the forgetting mechanisms differ for declarative and procedural knowledge. Based on the premises that procedural memory is more robust and much less vulnerable to decay, this theory suggests that long-term skill retention should benefit from developing sufficient procedural knowledge. As activations of declarative knowledge have to be strong enough for proceduralization (generating new production rules), massed practice with less spacing is hypothesized to be more effective in bringing learners' task knowledge from the transition stage to the final stage of learning.

2.2.3 Empirical research on spacing for cognitive skill acquisition and retention

Regarding the effects of temporal distribution on cognitive skill *acquisition*, the macro-level foreign language program evaluation studies reviewed in the previous section have generally supported that massed practice works better than distributed practice for skill acquisition or learning (Collins & White, 2011; Serrano, 2011; Serrano & Muñoz, 2007). The laboratory study by Suzuki (2017a), which focused on grammar learning for oral production, also found an advantage of massed practice over distributed practice in the accuracy measures starting from the beginning of the third training session. In addition, Beaunieux et al. (2006) showed the benefits of massed practice for the proceduralization of a cognitive procedural learning task, i.e., a Tower of Toronto task. Though not designed to test the interactions between spacing and skill acquisition, this study included two experiments with Experiment 1 conducted in a massed training condition (40 trials in one session), and Experiment 2 in a distributed training condition (4 sessions with 8 trials in each session; ISI was 1 day). Results showed that the massed group found the optimal solution to task (i.e., a minimum of 15 moves) in Trial 16, while the spaced group, at least on average, did not reach the optimal solution until in the last session, approximately in Trial 33. The authors pointed out that distributed learning "slows down the process of cognitive procedural learning" (Beaunieux et al., 2006, p. 521). This study, however, only looked at skill acquisition and not at retention.

When it comes to the effects of temporal distribution of practice on cognitive skill *retention,* as reviewed in the previous section, the superiority of distributed over massed practice has been predominantly found in educational psychology (e.g.,

Gluckman et al., 2014; Kapler et al., 2015; Rohrer & Taylor, 2006; Vlach & Sandhofer, 2012) and SLA (Bird, 2010; Miles, 2014; Rogers, 2015). It is true that these studies looked at higher-level learning; it should be noted, however, that these studies only focused on whether learners were able to retrieve the knowledge or rules learned and apply them in new contexts given enough time in offline tasks. The learning and retention outcomes were only measured in percentage accuracy, without taking cognitive fluency or automaticity into account. From the skill retention perspective, all these studies seem to have focused on the first stage of learning (i.e., the cognitive/declarative stage). From this perspective, these findings are actually in line with the prediction of skill retention theory, i.e., if learning stops at the declarative stage, distributed practice works better than massed for retention of skill performance. Suzuki and DeKeyser (2015), a study that investigated the effects of spacing on complex skill acquisition and used the RT measure for cognitive fluency or automaticity, demonstrated an advantage for the massed practice condition on speed measures for long-term retention (4 weeks delay) of skill performance in the oral picture completion task. Suzuki (2017a), as described in the earlier section, found an advantage for massed practice on the automatization of L2 morphology. As these two studies focused on the stage of proceduralization and the incipient stage of automatization, these findings seem to support the prediction of skill retention theory that massed practice works better than distributed practice in moving learners' task knowledge from the second stage to the third stage of learning.

Up till this point, to the best of my knowledge, the only study that was designed to directly test the hypothesis of learning schedules interacting with

knowledge types is Paik and Ritter (2016). This study examined how learning schedules (distributed, massed and hybrid schedules) interact with knowledge types (declarative, procedural and perceptual-motor). In this study, 40 participants were randomly assigned to four training conditions (10 per condition). Participants in each condition went through eight training sessions (30-min each), and in each training session participants engaged in three different types of training tasks. The eight training sessions all occurred within a timeframe of two weeks, four days per week from Monday to Thursday, were distributed in four different training schedules: the distributed group had eight training sessions in eight days, one session per day (1-1-1-1-1-1-1-1); the massed group had eight sessions in two consecutive days, 4 sessions per day (0-0-0-0-0-0-4-4); the hybrid-distributed group had eight sessions in six days, one session at the 1st, 3rd, 4th, 5th, and 7th day and 3 sessions at the 8th day (1-0-1-1-1-0-1-3); and the hybrid-massed group had eight sessions in four days over one week with some massed sessions (0-0-0-0-2-3-2-1). The retention tests were administered 3 weeks after the last training session of each group. Three different learning tasks were designed to evaluate the learning and retention of the three knowledge types: the declarative task was a Japanese-English vocabulary learning task, the procedural task was a Tower of Hanoi task, and the perceptual-motor task was an Inverted Pendulum task. The inverted pendulum task is a balancing task in which participants were asked to keep the pendulum (stick) vertical by tilting the device (iPod) at the bottom of the stick. Both accuracy and latency (RT) were used as outcome measures for the first two tasks; for the inverted pendulum task, the time that the pendulum can be held was used as the outcome measure. Results showed that there were no statistically

significant differences between the four training schedules for the declarative and procedural tasks for both learning and retention. For the perceptual-motor task, however, it was found that the hybrid-mass schedule led to statistically significant better learning and retention than the distributed schedule; descriptively, the hybrid-massed, massed, and hybrid-distributed schedules all led to better learning and retention than the distributed schedule. The authors concluded that the results lent support to the skill retention theory (Kim et al. 2013) in that the optimal spacing for the learning and retention depends on the knowledge type to be learned.

This study did not find an advantage for the distributed schedule over the massed one for the *retention* of declarative knowledge. This finding is inconsistent with the findings of many previous studies. It should be noted, however, that though the differences in accuracy rate between the distributed and massed groups on the retention test were not statistically significant, descriptively, the accuracy rate for the distributed group was much higher than for the massed group (51% vs 28%), with the rates for the hybrid groups somewhere in between (43% and 40%). The insignificant result may be partly due to the small sample size for each condition (10 per condition). Another important difference between this study and many previous studies is that this study included eight training sessions while the others typically only had two sessions. Future research is needed to see whether frequency of training sessions plays a role in mediating the effects of distributed practice. A final difference between this study and the previous studies is that the massed condition in this study is somewhat distributed, i.e., 4 sessions each on two consecutive days, rather than having 8 sessions on the same day, as in previous studies that usually have only a

single long session for the massed condition. This may have potentially contributed to the reduced effects of the distributed practice compared to the massed condition.

While there were noticeable losses in terms of both accuracy and RT in the declarative knowledge task from the last practice session to the retention session (RI = 3 weeks), performance in the last practice session was almost the same as in the retention session for the procedural task; in other words, the procedural knowledge did not decay over a retention interval of three weeks. This result suggests that procedural knowledge is much less susceptible to decay than declarative knowledge, providing empirical support to one of the most important premises of skill retention theory (Kim et al., 2013). Another noteworthy finding regarding procedural knowledge learning is that the four different training schedules did not have any significant effect on the learning and retention performance in the procedural task. The authors then suggested that "the most important factor for acquiring and retaining knowledge and skills in a task that requires procedural knowledge is not the training schedule, but simply the amount of practice" (p. 287). Due to the small sample size for each training condition, I think this conclusion cannot yet be seen as definitive. More empirical research is required on this regard.

The above studies have provided valuable information regarding the effects of spacing on skill acquisition and retention in general; no empirical studies, however, have systematically examined the interactions between temporal spacing of practice and *stages* of skill acquisition and retention. Such research would require a factorial design with levels of spacing for the different stages of skill acquisition. The primary difficulty in doing that lies in empirically pinpointing the boundaries of different

stages of learning. Due to variation of individual differences during skill acquisition, deciding the boundaries of the learning stages for each learner can almost be impossible to be done beforehand, and can only be done through post-hoc analysis. To tackle this difficulty, future research endeavors are absolutely needed.

The present study, as an exploratory study in this regard, does *not* attempt to *test* the interaction between spacing and stages of skill acquisition; instead, it attempts to document the course of learning in the (three) training sessions (which encompass the first two stages of learning and the incipient phase of the third stage of learning) under two spacing conditions (massed vs. distributed). Stages of learning in the present study are operationalized as the learning performance at the end of each training session, which does not necessarily correspond to the three stages of learning (declarative, transitional, procedural) from skill acquisition theory. This operationalization allows for the examination of the effects of treatment (i.e., spacing) and individual differences in cognitive aptitudes at different time points in time (after the same amounts of practice) during the learning processes.

2.3 L2 Mandarin tonal word learning

2.3.1 Mandarin tone

Mandarin Chinese is a tonal language that employs pitch variations to distinguish lexical meaning (Chao, 1948). Mandarin Chinese has four full lexical tones, in addition to a neutral tone that occurs only on weak unstressed syllables. The canonical forms of the four main tones are typically described in terms of pitch height and contour when they are in isolation: Tone 1 (T1) is called a high-level tone

(marked by the diacritic " ˉ " in Pinyin, the standard romanization of Mandarin Chinese), Tone 2 (T2) a high-rising tone (" ˊ "), Tone 3 (T3) a low-dipping tone (" ˇ "), and Tone 4 (T4) a high-falling tone (" ˋ ").

Chao (1930) introduced a system of describing the canonical patterns of these four tones based on a pitch scale over time. This system involved numeric notation, using five levels of pitch height to describe the change in pitch of a tone over time. These numbers represent points in a speaker's normal range of pitch that are equally spaced apart. The highest point of a person's pitch would be labeled 5, and the lowest would be labeled 1. Using Chao's notation, T1 is labeled as [55], which means that over the course of the time in which the tone being produced, it stays at a high pitch. T2 is labeled as [35], starting in the middle range of one's pitch then rising. T3 is traditionally labeled as [214], but it is now more widely accepted as [212] among Chinese linguists (Yang, 2016) based on acoustic analysis. T3 starts low, then dips to the bottom of one's range and then rises a bit, but still within the lower register of one's voice. Finally, T4 is traditionally labeled [51], showing a sharp fall from the highest pitch to the lowest pitch; however, acoustic analysis shows that this tone should be more accurately labeled as [53], as it does not actually fall to the bottom to one's range (Duanmu, 2007).

Tones may undergo alternations depending on their tonal contexts. The variation of tonal patterns depending on the surrounding tonal contexts in connected speech is called *tone sandhi* (Chen, 2000). T3 has three allotonic variants or alternations. Most often, T3 is just a *low* tone [21], or "Half-T3" in Zhang's (2016) term. It is just a low tone when it is followed by T1, T2, and T4, and can be just a low

tone at the utterance-final position when it is not emphasized. T3 is pronounced as a *low-dipping* tone [212], or "Full-T3" in Zhang's (2016) term, only when it is in isolation or at the utterance-final position for emphasis. When T3 is followed by another T3, the first T3 is pronounced as a *rising* tone [35] (i.e., T2), or "Raised-T3" in Zhang's (2016) term, which is known as the T3 sandhi.

In theoretical Chinese linguistics, it is still under debate whether the low-dipping [212] variant or the low [21] variant should be the underlying form of T3 (for details see Zhang, 2016). Among researchers and instructors in Chinese as a second language, however, it is now widely accepted that T3 should be best taught just as a low tone (Yang, 2016; Zhang, 2016). For the purpose of the present study focusing on oral production of disyllabic words, T3 is introduced as a low tone, with the low-dipping variant introduced as a special case when it is in isolation or for emphasis, and with the rising variant introduced as a result of a phonological process, i.e., when it precedes another T3 (see Appendices D and G for the instruction sheets presented to participants).

2.3.2 Research on L2 tonal word learning by NSs of non-tonal languages

Successful learning of Mandarin words entails successful learning (in terms of both perception and production) of the tones and the ability to use the tonal contrasts to distinguish meaning in lexical access, an aspect of language processing that NSs of non-tonal languages have never attended to before in their L1 use. Earlier L2 tone learning studies primarily focused on the lower-level speech sound learning (e.g., Leather, 1990; Wang, Jongman, & Sereno, 2003; Wang, Spence, Jongman, & Sereno, 1999), without considering tone learning as an integrative part of word learning. It is

a relatively recent development that researchers started to investigate tone learning at the lexical level, i.e., tonal word learning (e.g., Bowles, Chang, & Karuzis, 2016; Chandrasekaran, Sampath, & Wong, 2010; Chang & Bowles, 2015; Cooper & Wang, 2012, 2013; M. Li & DeKeyser, in press; Perrachione, Lee, Ha, & Wong, 2011; Wong & Perrachione, 2007).

The training studies on L2 learning of tonal vocabulary by NSs of nontonal languages have attempted to identify factors (individual or contextual) that may affect their learning success. Wong and Perrachione (2007) found that native English speakers who had no prior experience with any tone language were capable of learning to use pitch patterns for lexical identification, although large variability exists in learning success between individuals. In addition, they found that learning success was predicted by the ability to identify pitch patterns in a nonlexical context, which was associated with musical experience. Chandrasekaran et al. (2010) further determined that individual variability in lower-level phonetic cue weighting, particularly pitch direction identification, contributed to differential success in lexical tone-word identification. However, cognitive measures in phonological awareness and verbal working memory as measured by a set of subtests from the *Woodcock-Johnson Tests of Cognitive Abilities* (Woodcock, 1997), i.e., Sound Bending, Numbers Reversed, and Auditory Working Memory, did not distinguish good and poor learners of tone-word identification.

Cooper and Wang (2012) investigated the interactive effects of linguistic and musical experience on non-native tone perception and word learning. In their study, four groups of participants, differing on L1 background (tonal versus nontonal) and

musical background, i.e., native Thai musicians, native Thai non-musicians, native English musicians and native English non-musicians, engaged in Cantonese tone-word learning. Their results showed that (a) musical experience was more advantageous than a tone language background for both tone identification and tone word identification, (b) musical training did not add much influence on tone-word identification for those whose L1 is tonal, and (c) pre-training tone identification and musical aptitude scores positively predicted tone-word learning success for English listeners but not for Thai listeners. Cooper and Wang (2013) examined the effects of lower-level perceptual tone training on the higher-level tone-word learning. They found that English non-musicians, who received three sessions (30 minutes each) of auditory perceptual tone training before engaging in tone-word identification training, obtained a similar level of proficiency in tone-word identification as musicians, and performed significantly better than non-musicians who had no tone training. They concluded that lower-level perceptual ability enhanced by short-term tone training significantly contributed to the ability to use tonal contrasts to distinguish word meaning, and highlighted the role of bottom-up processes in speech perception and higher-level linguistic learning.

Bowles et al. (2016) presents perhaps the most comprehensive examination of a wide range of aptitudes on perceptive L2 tonal word learning. In this study, 160 NSs of English with no previous tone language experience completed a Mandarin word learning task over six training sessions and a battery of cognitive tests. The Mandarin word learning task consisted of the learning of the sound-meaning mappings of 24 Mandarin pseudowords ("pseudowords" in the sense that the sound-

meaning mappings are illegal in real language use). Eight of the 24 tonal word forms were monosyllabic and the other 16 disyllabic. The 8 monosyllabic word forms consist of two monosyllabic minimal tonal quadruplets; the 16 disyllabic word forms consist of four disyllabic minimal tonal quadruplets with tonal contrast on either the first or the final syllable. The battery of the cognitive tests included four pitch ability tests (two linguistic, i.e., Tone Discrimination and Tone Identification based on Mandarin tones, and two nonlinguistic, i.e., Pitch STM and Pitch contour identification using sine waves), two measures of musicality (i.e., musical aptitudes and musical experience), six foreign language aptitude tests (i.e., Consonant Discrimination, Nonword Span for phonological STM, Running Memory Span for the updating function of working memory, Antisaccade Analogue for the executive function of inhibition, Serial Reaction Time for implicit induction, and Paired Associates for verbal rote learning), and two general cognitive ability tests (i.e., the Wonderlic Contemporary Cognitive Ability Test and the Letter Sets Test). The tonal word learning outcome measures were the Penultimate Accuracy at the fifth test phase with trained stimuli, and the Final Accuracy at the sixth test phase with stimuli from novel talkers. Results showed that the penultimate accuracy was significantly predicted by 6 cognitive measures: the two linguistic pitch ability tests (including tone identification and discrimination), nonword span, paired associates, months of private music lessons and the letter sets. The final accuracy with stimuli from new talkers were significantly predicted by pitch contour identification and pitch STM (the two nonlinguistic pitch ability measures), in addition to the same 6 cognitive predictors for the penultimate accuracy. The findings demonstrated that pitch ability

provided additional predictive power for L2 tonal word learning beyond musicality, general L2 aptitudes and general cognitive abilities.

Drawing from the same data set for Bowles et al. (2016), instead of examining individual differences in cognitive variables, Chang and Bowles (2015) focused on investigating the effects of an external contextual variable, i.e., contextual *phonetic* variability, on perceptive L2 tonal word learning. Contextual phonetic variability refers to variability caused by coarticulatory modification that is driven by articulatory influence from nearby sounds (or tones in this context). It was found that "tones were acquired less successfully in disyllables than in monosyllables, and the relative difficulty of disyllables was closely related to contextual tonal variability" (p. 3703). For tone learning in disyllabic contexts, phonetic tonal variability was inversely related to learning success, that is, larger tonal variability resulted in less successful learning of the tones.

Perrachione et al. (2011) is the only study so far that has examined the interaction between individual difference variables and an external contextual variable (e.g., input variability) for L2 tonal word learning. In Perrachione et al. (2011), input variability refers to talker variability: in the low-variability training condition, the stimuli were from only one talker, while in the high-variability training condition, the stimuli were from four talkers. Perrachione et al. (2011) examined how learners of different levels of pitch perceptual abilities fared with the above two training conditions that varied in talker/input variability. They found that "high-variability training enhanced learning only for individuals with strong perceptual abilities", while "learners with weaker perceptual abilities were actually impaired by

high-variability training relative to a low-variability condition" (p. 461). In other words, an aptitude-treatment interaction (ATI) is demonstrated in this study, which highlights the importance of considering individual differences in learners' preexisting aptitude when evaluating the efficacy of instructional interventions.

While all the above-described studies have only focused on perceptive tonal word learning, i.e., tonal word identification (in the form of sound-meaning mappings), Li and DeKeyser (in press) is the only study so far that has included production training for L2 tonal word learning. Their study was designed to test the skill-specificity hypothesis in a study on the L2 learning of Mandarin tonal words (monosyllabic) by naïve adults English NSs. Participants went through either perception practice or production practice over three training sessions and were administered both perceptive and productive tests by the end of the last training session. The results showed that students' performance was far worse when tested in the reverse skill than when tested on the practiced skill in terms of both accuracy and response times, providing strong support to the skill specificity of practice effects. Musical ability was also found to facilitate tonal word learning in both perception and production.

There are various gaps, then, in research on L2 tonal word learning. Most studies have focused on perceptive tonal word learning, but it is also very important to study oral production learning of tonal words by NSs of nontonal languages, and more empirical research is absolutely needed. In addition, the timing seems to be ripe to move on to investigate L2 learning of disyllabic tonal words because Chang and Bowles (2015) have shown convincing evidence indicating "limited relevance of

monosyllable-based data on Mandarin learning for the disyllabic majority of the Mandarin lexicon" (p. 3703). To the best of my knowledge, no studies have looked at *oral production* training of *disyllabic* tonal words in adult L2 learning. In addition, no studies have attempted to identify a relatively comprehensive list of aptitudes for L2 learning of oral *production* of tonal words. Finally, the temporal distribution issue has not been examined with L2 learning of oral production of tonal words. These are the gaps the present dissertation attempts to fill.

2.4 Cognitive aptitudes for L2 tonal word learning

In addition to temporal distribution of practice, a second goal of this present dissertation is to explore the cognitive processes underlying L2 learning and retention of oral production of Mandarin words by NSs of nontonal languages under different practice distribution conditions. To investigate such research questions, apart from brain imaging studies (e.g., Hubert et al., 2007), promising behavioral approaches used in the literature include (a) studying the correlations between cognitive ability measures and skill performance levels during the course of learning (e.g., Ackerman, 1988; Ackerman & Cianciolo, 2000; Beaunieux et al., 2006), and (b) studying the interactions of cognitive aptitudes with the treatments (see DeKeyser 2012). The present study attempts to combine the two behavioral approaches, i.e., through correlation and/or interaction analyses of outcome performance at different stages (i.e., performance at the end of each training session) of learning as a function of aptitude measures and types of treatment (i.e., different practice distributions). To unravel the underlying cognitive learning processes, the key is then to identify cognitive aptitudes that may play a role during the learning of the target skills.

Let us first do a task analysis of the target skill to be learned, i.e., L2 learning of oral production of Mandarin disyllabic words by native English-speaking adults. From a skill acquisition perspective, fluent L2 oral production of Mandarin disyllabic words requires the mastery of two types of knowledge, i.e., declarative knowledge of word meanings (or spelling-meaning mappings) and procedural knowledge in oral production (of the segments and tones), which can also be considered two sub-components of oral Mandarin word production. While the declarative component of word knowledge (i.e., spelling-meaning mappings) remains declarative throughout the whole learning process, the acquisition of the knowledge required for oral production of Mandarin disyllabic words goes through the typical three stages of skill acquisition, i.e., from declarative, to proceduralization, and then automatization, at least in the context of this study. As the focus of the present study is *tone* production of the disyllabic words, the learning difficulty of the segments that constitute the target words is reduced to minimum (described in Section 5.4.1) by choosing segmental phonemes that are easy for English NSs to pronounce. The learning task for the procedural part is then mainly Mandarin tone sequence production. To develop skills in Mandarin tone sequence production (in disyllabic words), students first need to learn the following components of declarative knowledge about Mandarin tone: the role of tones in differentiating lexical meaning in Mandarin, the properties of each of the four Mandarin tones (i.e., pitch contours), the mappings of the verbal descriptions of the tones (high, rising, low, falling) and the tone marks (- ˊ ˇ ˋ), as Pinyin (visual symbols) will be used to help learning, and the rules about tone changes or non-changes in disyllabic words. The first three components help guide

the learning of oral production in monosyllables, and the fourth component consists of the phonological rules students need to learn to apply in oral disyllabic word production. To facilitate learning, students will also hear the training words and target words, which involves auditory perceptual learning.

For disyllabic words that do not involve T3, e.g., "yīnghuā," oral production involves the direct application of the description of the canonical tones in oral production. For words that involve two T3s, e.g., "gǎnlǎn", students need to be able to apply the phonological T3 sandhi rule, (i.e., the first T3 has to be changed to a rising tone when it precedes another T3) when orally producing it, and keep in mind that the second T3 is just a low tone (without the rising tail). When orally producing words with one T3, e.g., "kǔguā", students need to bear in mind that T3 is just low, without the rising tail. The learning of these words is aided by hearing the auditory pronunciation of each word when it is visually presented (including its Pinyin with tone marks, and its meaning illustrated in pictures together with English translation).

The learning task of oral production of Mandarin disyllabic words is indeed a complex and difficult task for English native speakers who are not used to orally producing tones in the four particular ways, nor to using tonal contrasts to distinguish lexical meaning. This task involves declarative learning of meaning-spelling mappings and the above-listed components of tone knowledge, and procedural learning of oral production of the tones in disyllabic sequences, including rule learning (i.e., the phonological T3 Sandhi rule), the conversion of the visual T3 mark " ˇ " to a verbal "low" tone, in addition to the oral production (and auditory perception) learning of each of the canonical tones. As a cognitive speech-motor skill

acquisition task, L2 learning of oral production of Mandarin disyllabic words by adult English NSs seems to involve not only cognitive processes in working memory (including both short-term storage and processing), but also learning in the declarative and procedural memory systems (Ullman, 2015), and interactions between these memory systems. In the following, I explain why and how these memory systems were expected to be involved in this learning task, and what aptitudes were expected to play a role during which learning processes.

2.4.1 Working memory

Working memory (WM), the limited capacity to temporarily store and manipulate information in immediate memory, is expected to play a crucial role in this L2 tonal word learning task in oral production, especially at the initial stages. From a skill acquisition perspective, the initial stage of cognitive skill acquisition is largely cognitive (Anderson, 2000), and there has been empirical evidence that the first stage of skill acquisition is associated with general intelligence (Ackerman, 1988) and working memory (Beaunieux et al., 2006). Different models of WM differ in how they conceptualize WM capacity; however, all models generally agree that WM is a limited-capacity system that regulates the processing, storage and retrieval of temporary information (e.g., Baddeley, 2012; Engle, 2002). In the present study, I adopt Baddeley's seminal multicomponent model of working memory (Baddeley & Hitch, 1974; Baddeley, 2000, 2003, 2012), which remains dominant in the field of language learning. In the original model (Baddeley & Hitch, 1974), the WM system contains a domain-general attentional control system, the central executive, aided by two domain-specific storage-based slave systems, the phonological loop, specialized

for briefly storing and maintaining verbal and acoustic information, and the visuo-spatial sketchpad, responsible for visual and spatial information. A more recent addition is the episodic buffer, which is "a limited capacity system that provides temporal storage of information held in multimodal code, which is capable of binding information from the subsidiary systems, and from long-term memory, into a unitary episodic representation" (Baddeley, 2000, p. 417). The central executive is responsible for controlling attentional recourse to the slave systems and information from long-term memory (Baddeley, 2000), and performs a range of attentional functions, such as updating, inhibiting, and switching (Miyake & Friedman, 2012). See Figure 3 for the updated version of this model presented in Baddeley (2000).

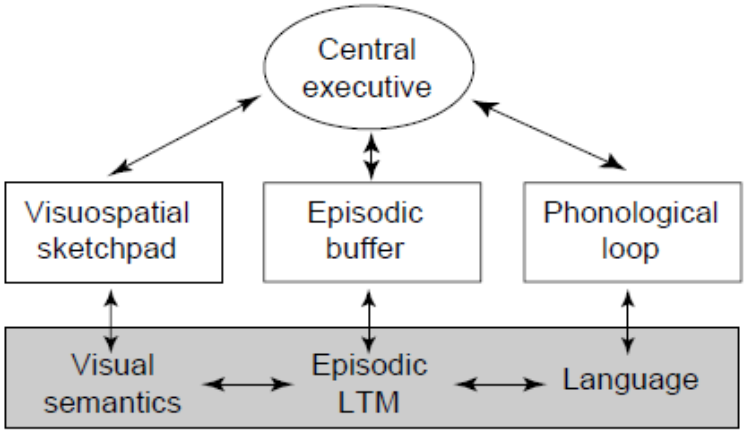


Figure 3. The updated version of the multi-component working memory model, from Baddeley (2000), p. 421. Shaded areas represent crystallized systems and unshaded areas represent fluid systems.

A large body of empirical evidence has shown that the phonological loop plays a crucial role in language learning. In fact, the phonological loop has been termed the "language learning device" because it plays a critical role for the learning

of "novel phonological forms of new words" (Baddeley, Gathercole, & Papagno, 1998). Phonological short-term memory (PSTM), the storage component of the phonological loop, facilitates language learning by providing temporal storage of novel phonological traces until more permanent representations can be formed. PSTM has typically been measured through simple span tasks that require the storage of verbal units, such as random digits, letters, words, or nonwords (Williams, 2012). For L2 learning, empirical evidence has shown that PSTM contributes to not only L2 vocabulary learning in children (e.g., Cheung, 1996; Masoura & Gathercole, 2005) and in adults (e.g., P. W. B. Atkins & Baddeley, 1998; Baddeley et al., 1998; Baddeley, Papagno, & Vallar, 1988; Kaushanskaya, 2012; Papagno, Valentine, & Baddeley, 1991) but also to L2 grammar (morphosyntax) learning independent of L2 vocabulary in children (e.g., French & O'Brien, 2008; Verhagen & Leseman, 2016; Verhagen, Leseman, & Messer, 2015) and adults (Martin & Ellis, 2012). It is worth noting that few of the above studies on L2 adult learning have tested vocabulary or grammar performance using oral productive tasks, except Martin and Ellis (2012). Studies on L2 oral fluency development have identified PSTM as a predictor of L2 oral fluency development (as measured by temporal/hesitation measures) (e.g., O'Brien, Segalowitz, Freed, & Collentine, 2007). When it comes to L2 Chinese spoken word learning, Wei (2015) found that PSTM independently predicted Chinese spoken word learning (including 1-syllable, 2-syllable, and 3-syllable words) in native English-speaking children (4^{th} grade, 10-11 years old). In addition, in Bowles et al. (2016), PSTM (as measured by a nonword span task) was found a significant predictor of tonal word learning success in sound-word identification accuracy. Based

on the above evidence, it is reasonable to expect PSTM to play a role in L2 learning of oral production of Mandarin disyllabic words by English adult learners.

Verbal *working memory*, as compared to PSTM, emphasizes the combined *processing* and storage of verbal information, or the executive control component of WM when processing verbal information. The executive attentional control functions of WM have been implicated in a wide range of complex cognitive tasks, such as reasoning, learning, problem solving, abstraction, and comprehension (Linck, Osthus, Koeth, & Bunting, 2014). For a cognitive speech-motor task in L2 oral production of Mandarin disyllabic words that involves the learning of a phonological rule, the executive control function is expected to be taxed, especially at the initial learning stage, because learners need to keep the phonological rules in mind while applying one of them in the appropriate context in real-time oral production of the word (integrating tones in syllables). In the L2 learning literature, the complex verbal WM has been typically measured through complex span tasks, such as listening span (e.g., Martin & Ellis, 2012; Sanz, Lin, Lado, Stafford, & Bowden, 2014; Verhagen & Leseman, 2016; Wright, 2013), reading span (e.g., Sanz et al., 2014), and operation span (e.g., Linck & Weiss, 2011; Serafini & Sanz, 2015; Suzuki & DeKeyser, 2016), which require both storage and processing or manipulation functions. The verbal complex WM has been found to play a role in various aspects of L2 learning and development, such as reading comprehension (e.g., Leeser, 2007), L2 grammar learning as measured in oral production tasks (e.g., Martin & Ellis, 2012; Serafini & Sanz, 2015; Suzuki & DeKeyser, 2016; Wright, 2013) and receptive tasks, such as GJT (e.g., Robinson, 2002; Serafini & Sanz, 2015), oral fluency at the discourse level

(e.g., O'Brien et al., 2007), and general proficiency, such as reading and listening (e.g., Kormos & Sáfár, 2008; Linck et al., 2013). The complex verbal WM has also been found to interact with different types of instruction or treatment (e.g., corrective feedback), more explicit or implicit (e.g., Mackey, Philp, Egi, Fujii, & Tatsumi, 2002; Sanz et al., 2014; Yilmaz, 2013) and different practice distribution conditions (Suzuki & DeKeyser, 2016).

In fact, far fewer studies have attempted to investigate the role of complex verbal WM in L2 vocabulary or word learning. Verhagen and Leseman (2016) investigated how verbal STM and WM relate to *child* L2 acquisition of vocabulary and grammar in a naturalistic setting and child L1 acquisition. They worked with a group of Turkish children who were learning Dutch as an L2 and a group of Dutch monolingual children; the mean age for both groups was about 5 years. Word recall, Dutch-like nonword recall, and Dutch-unlike nonword recall were used as measures of verbal STM; backward digit recall and listening recall were used to measure verbal WM capacity. Vocabulary was tested in a receptive vocab task, and grammar in both comprehension and production tasks. Results showed that verbal STM and verbal WM were identified as two separate latent factors in both groups. Verbal STM significantly predicted both vocabulary and grammar performance, but verbal WM predicted only grammar performance but not vocabulary performance, in both groups. Note that only receptive vocabulary was tested, which might have contributed to the inconsistent finding with the following study.

Martin & Ellis (2012) investigated the roles of PSTM and WM in *adult* L2 learning of vocabulary and grammar in an artificial foreign language. PSTM was

measured by two tasks, nonword recognition and nonword repetition, and WM was measured by a listening span task. Participants first learned the singular forms of the vocabulary before being exposed to the word order and plural forms in sentence contexts, without explicit instruction. After two training sessions, participants were tested on their ability to induce the plural forms and to generalize the forms to novel sentences in a production task and a comprehension task. Through hierarchical multiple regression analyses, the results demonstrated that PSTM (nonword repetition) and WM each had significant independent effects on L2 vocabulary learning and on L2 grammar learning.

Kapa and Colombo (2014) investigated whether executive function abilities in adults and preschool children (age 4 to 6 years old) predict their success in learning a novel artificial language. Executive function was defined following Miyake, Friedman, Emerson, Witzki, and Howerter (2000), and was measured by using three tasks, i.e., a visual Simon task, a flanker task, each of which was used to test attentional monitoring (which is part of the process of updating) and inhibitory control, and a Wisconsin card sorting test that was used to test the shifting function. The artificial language contains 12 nouns that were animate, real-world objects, either animals or human, and 4 verbs that described motions. The sentence word order was verb - noun 1 (agent) – noun 2 (theme). After two training session of hourly long for each, learning outcomes were measured in a number of tasks, including productive vocab tasks (picture naming), a receptive vocab task (sound-picture mappings), a productive sentence task (sentence narration from video), a receptive sentence task (identifying the video from two choices correctly describing the sentence), and a GJT.

Principle components analysis of the outcome measures identified only one single factor, which was interpreted as vocabulary knowledge in all the outcome tasks due to the simplicity of the language. Results from multiple regression analyses showed that after controlling for L1 receptive vocabulary knowledge and WM ability (as measured by simple digit span tasks, forward and backward), adults' artificial language performance was predicted by inhibitory control ability, and children's artificial language performance was predicted by attentional monitoring and shifting abilities. These results suggest that "EF processes may be employed during initial stages of language learning, particularly vocabulary acquisition" (Kapa & Colombo, 2014, p. 237).

Although verbal WM (complex) was not found to predict receptive vocabulary in Verhagen and Leseman (2016), the findings from Martin and Ellis (2012) and Kapa and Colombo (2014) provided evidence that the executive function component of WM is likely to play a role in L2 word learning, especially when oral production of L2 words was also taken into account in addition to receptive L2 knowledge. It is tempting to suggest that oral word production learning in a novel language seems to rely more on the executive function of WM than receptive learning of new words.

From a developmental perspective, researchers have started to examine whether the role of PSTM and/or complex WM in L2 learning (with adults) may depend on the proficiency level of the learners. Using a cross-sectional design, Hummel (2009) found that PSTM was a significant predictor of L2 proficiency performance (in vocab, grammar, and reading comprehension) for the lower

proficiency group, but it did not turn out to be a significant predictor for the higher proficiency group. Using a longitudinal design, that is, by tracking L2 learning through a period of a semester, Serafini and Sanz (2015) found that significant correlations between PSTM (as measured by digit span) and L2 performance emerged only for lower proficiency learners, but not for higher proficiency learners. The same pattern was also observed for WM (as measured by Ospan), but to a lesser extent. These findings seem to suggest a decreasing impact of cognitive ability in L2 learning with increasing L2 proficiency. Linck et al. (2013), in a study that aimed to identify potential cognitive predictors for advanced L2 proficiency attainment, found that working memory, i.e., PSTM and the executive function of task switching, was a good predictor for predicting high-level attainment. The findings from these studies are informative, but a more fruitful line of research on this may be to track the roles of WM (including PSTM) in L2 learning of a particular skill at different points in time, which is what this present study attempts to do.

I will now turn to visuospatial WM, the subsystem that serves to manipulate and temporarily maintain visual and spatial information (Baddeley, 2003). Though visuospatial WM is considered to be of less relevance to language learning than the verbal WM system, including PSTM, it was suggested that it may play a role in reading comprehension (Baddeley, 2003), and Chinese character learning (Opitz, Schneiders, Krick, & Mecklinger, 2014). In the present study, visual forms of the words in Pinyin are presented to help aid the formation of more stable long-term phonological representation of the tonal words. In a word naming task, i.e., when students are presented with a tonal word in Pinyin, students need to process the visual

information of the two tonal marks in the disyllabic sequence, and pronounce the tone sequence imposed on the two syllables accordingly. For words that involve a T3, e.g., “kǔguā”, some mental manipulation of the T3 tone mark “ ˇ ” seems necessary to convert the visual T3 mark “ ˇ ” to a “low” tone. When presented with words that involve two T3s, e.g., “gǎnlǎn”, more visual manipulation processes seem necessary when applying the T3 Sandhi rule, i.e., the first T3 should be changed to a rising tone, while still keeping in mind that the second T3 should be pronounced as low. As these processing requires not only brief storage but also manipulation of visual and spatial information of the tonal marks, I expect that visuospatial WM may play a role in the oral tonal word production task.

In an attempt to identify specific aptitudes for tone learning, Bowles et al. (2016) found that pitch STM, the ability to hold nonlinguistic tonal information in STM in the face of intervening tones, was a significant predictor of receptive Mandarin tonal word learning when tested on transfer to new talkers. It is therefore reasonable to hypothesize that pitch STM may also play a facilitative role in productive tonal word learning.

2.4.2 Declarative memory and Procedural memory

Declarative memory (DM) and procedural memory (PM) are considered the two most important long-term memory systems in the human brain due to the wide range of functions and domains they subserve (Ullman, 2015). The functions of the DM system are to learn “information about facts (semantic knowledge) and events (episodic memory)” (Ullman, 2015, p. 137). The PM system, less well understood, is considered to be “the nondeclarative system that underlies both motor and cognitive

skill and habit learning" (Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter, & Wong, 2014). As the learning task in the present study, i.e., L2 learning of oral production of disyllabic Mandarin words by English NSs, is a cognitive speech-motor skill learning task, that involves not only declarative learning of meaning-spelling mappings of the target words and knowledge about the tones, but also procedural learning of oral production of the tones in disyllabic sequences (including the learning of a phonological rule), it is reasonable to expect that both DM and PM should be involved during the learning process of achieving proficiency in this task.

The Declarative/Procedural model (Ullman, 2001, 2004, 2015), when it comes to L2 learning, predicts that DM should be responsible for learning all "idiosyncratic knowledge" in the language, such as word forms, meanings and mappings between them. Due to the flexible nature of DM, DM is also expected to be able to learn grammatical rules in the form of declarative knowledge (i.e., as information about the rules). Procedural memory is expected to "underlie the learning and processing of sequences and rules" and to play an important role in grammar learning, which "should hold across linguistic subdomains, including syntax, morphology, and phonology" (Ullman, 2015, p. 141). Ullman also notes that while knowledge in DM can be learned rather rapidly, linguistic knowledge in the DM is expected to be learned "gradually" since PM learns from "repeated exposure" (p. 141). When it comes to L2 grammar learning that can involve both DM and PM, Ullman predicts that:

> "aspects of grammar should initially be learned in declarative memory. In parallel, procedural memory should also gradually learn

> grammatical knowledge. After sufficient experience with the language, procedural memory-based grammatical processing should tend to take precedence over analogous declarative knowledge, resulting in increasing automatization of the grammar." (Ullman, 2015, p. 143)

This hypothesis has received some support from behavioral studies that correlated declarative memory ability and procedural memory ability at different stages of learning (Hamrick, 2015; Morgan-Short et al., 2014). Morgan-Short and colleagues (2014) tested this hypothesis in an artificial language learning studies with adult learners. The target structure was the syntactic word order. Participants went through four language training sessions on four separate days. Syntactic development was assessed by using an auditory GJT at both early (after the 1st training session) and later stages (after the fourth training session) of learning. Declarative memory ability was assessed by using MLAT-Verbal, and the Continuous Visual Memory Task (CVMT), and procedural memory ability was measured by using a Weather Prediction Task (probabilistic), and a Tower of London Task (cognitive skill learning). The results demonstrated syntactic performance at the early stage was positively correlated to declarative memory ability, with no significant correlation with procedural memory ability. Syntactic performance at the late stage of learning, on the other hand, positively correlated with procedural learning ability, but not with declarative memory ability.

Hamrick (2015) investigated the role of individual differences in declarative and procedural memory abilities in the learning and retention of word order of a semi-artificial language under an incidental condition (with exposure only for about 20

min). Declarative memory ability was measured by using LLAMA-B, which is a picture-nonword association task, and procedural memory ability was measured by using a Serial Reaction Time (SRT) task. A surprise recognition test was immediately administered after exposure, and an identical surprise post-test after 1 to 3 weeks. It was found that “declarative memory abilities predicted performance on the immediate, but not delayed, recognition task, whereas procedural memory abilities predicted performance on the delayed, but not immediate, recognition task". In other words, participants relied on declarative memory to perform on the immediate surprise recognition test, but the reliance shifted to procedural memory in the delayed posttest.

These two studies seem to have provided good support to the hypothesis regarding the interaction between declarative memory system and procedural memory system with respective to L2 grammar learning. It is reasonable to expect that this relationship would hold for L2 learning of phonological rules. Therefore, I expect declarative memory ability to play a larger role at the early stage of L2 Mandarin tone word learning, and procedural memory ability to play a larger role at the later stages of learning and in retention after a period of inactivity.

Chapter 3: Purpose of the current study

The present dissertation research was motivated by three major research gaps in the body of literature on the issue of temporal distribution of practice in L2 learning. First, few studies have looked at the learning of a complex L2 skill in a laboratory setting with extraneous factors (e.g., prior knowledge, outside practice) well controlled. Among those that looked at the acquisition of a complex L2 skill, which typically involves memorization of declarative knowledge and skill acquisition, no previous research has done an elaborate task component analysis of the target skill and examined how the effects of ISI and RI might differ on the retention of the different subcomponents or types of knowledge or skill (i.e., declarative knowledge vs. skill). Second, in selecting the linguistic domain of the target skill, all those studies on L2 vocabulary learning have focused on the memorization of paired associates in the visual domain without taking auditory perception or oral production into account, while the few studies that looked at oral production skill learning only dealt with morphosyntactic rules, and none with the learning of phonology in oral production. Finally, few studies have investigated the roles of cognitive aptitudes at different stages of L2 learning and how the roles might differ depending on practice distribution or retention intervals.

In order to address the first two gaps, the present study examines the effects of temporal distribution of practice (relatively massed vs. distributed) on the automatization and retention of L2 Mandarin word production by a group of naïve native speakers of English. The target skill to be learned in the study was oral production of Mandarin disyllabic words from conceptualization to articulation. This

task is a complex skill that involves several subcomponents, e.g., declarative knowledge of meaning-word mappings (DK1); declarative knowledge about how to pronounce the Mandarin tones, including the phonological rule in disyllabic words (DK2); and procedural knowledge for articulation of the words (PK). While DK1 (i.e., meaning-word mappings) remains declarative throughout the whole learning process, DK2, the knowledge about how to pronounce Mandarin tones that is required for the acquisition of the oral tone production skill, goes through three stages of skill acquisition in the context of this study, from declarative, to proceduralized and (partially) automatized. DK2 might not be required when this knowledge is fully automatized.

Three outcome tasks were designed to measure the complex skill and the two major subcomponents: (a) declarative knowledge of meaning-word associations (DK1) and (b) oral tone production skill (encompassing DK2 and PK). An oral picture-naming task was designed to measure the global skill in oral Mandarin word production from meaning to articulation. A written picture-naming task was designed to measure DK1, i.e., declarative knowledge of meaning-word associations. An oral word-naming task (i.e., reading aloud from Pinyin) was designed to assess tone production skill, which definitely involves PK in oral articulation and may or may not involve DK2 (declarative knowledge about tones) depending on the extent to which DK2 is automatized. If DK2 is fully automatized, participants should be able to perform the task without the involvement of DK2 when prompted by Pinyin with tone marks; if not fully automatized, the part of DK2 that is not automatized is required to complete the task. The oral picture-naming task requires DK1 and oral tone

production skill, which requires PK (along with DK2 till the corresponding PK is fully automatized).

The selection of the target skill and the design of the three outcome tasks which tap into different subcomponents or different combinations of subcomponents allow us to explore whether the effects of ISI and RI differ depending on the type of knowledge to be acquired or retained, i.e., declarative knowledge only (in written picture naming) vs. oral production skills that typically involve a combination of declarative knowledge and procedural knowledge (in oral word naming or oral picture naming). In addition, for the oral *skills*, as oral picture naming requires a larger proportion of declarative knowledge to complete the task (DK1 plus PK and possibly DK2) than the oral word naming task does (PK and possibly DK2), a comparison of the effects of ISI and RI on these two outcome tasks allows us to see whether the effects differ depending on the proportion of declarative knowledge required to complete the oral production task. Furthermore, regarding the retention of tone production skill as measured by oral word naming, both practiced/old words and new words were included. A comparison of the effects of ISI and RI on oral word naming for old words versus new words allows us to see whether the temporal effects on retention of tone production skill differ in practiced linguistic contexts versus new contexts.

To address the third gap, this study also attempts to scrutinize the roles of cognitive aptitudes (including complex WM, PSTM, declarative memory ability, procedural memory ability, and musical aptitude) at different stages of L2 learning of

Mandarin tonal word production and in the retention of the learned knowledge or skills under different practice distribution conditions or retention intervals.

The ultimate goal of this study is to shed some light on the underlying mechanisms of learning and forgetting/retention of different types of knowledge and skills (that involve a mix of knowledge types) in distributed practice conditions with varying levels of spacing.

The present study also attempts to treat RI as a between-subjects variable, instead of a within-subjects variable, to avoid the confound introduced by testing at the shorter RI on retention performance at the longer RI. In addition, the present study uses both accuracy and RT measures for performance on the outcome tasks, except for the written picture naming task for which only the accuracy measure is used. The present study also documents the development of learning across the training sessions, in addition to assessing retention at the delayed posttest after an RI. Participants went through three training sessions (TS) in the study. A pretest in the form of oral word naming was administered at the beginning of the first training session (TS1). Starting from the end of TS1 to the end of the last training session (TS3), a pre- or post- session quiz in the form of an oral picture naming task was administered at the beginning and the end of each TS to keep track of learning and forgetting. For the delayed posttest (or the retention test), three tasks were administered – an oral picture naming task, a written picture-naming task, and an oral word naming task including practiced/old words and new words.

Chapter 4: Research Questions and Hypotheses

Four main research questions were put forward; sub- research questions were put forth when needed. Specific measurable hypotheses were formulated to examine each of the research questions in detail. The research questions and hypotheses are as follows.

RQ1: What are the effects of ISI on L2 learning of oral Mandarin tonal word production across training sessions?

- Hypothesis 1a: As all participants started from zero in learning Mandarin tonal word production (no participants had any prior knowledge about Mandarin tones or any of the target words), as this study has an experimental design with participants randomly assigned to experimental conditions, and as ISI can only start to exert an effect after an ISI has happened, a comparison of different ISI groups on performances in the pre- and post-session quizzes of TS2 and TS3 can demonstrate the effects of ISI on L2 learning (higher outcome performances, more learning/improvement). As massed practice would result in less forgetting between the training sessions, and is more likely to enable participants to move from the second to the third stage of skill acquisition, the massed practice group (ISI-1day) is expected to outperform the distributed practice group (ISI-1week) on oral picture naming accuracy, from TS2 pre-session quiz, TS2 post-session quiz, to TS3 pre-session quiz, and TS3 post-session quiz.
- Hypothesis 1b: The massed practice group is expected to respond faster than the distributed practice group in oral picture naming at the end of TS3, as

increased/deeper procedural knowledge is likely to translate into faster RTs. The RTs at the earlier stages (before the end of TS3) were not used because at those earlier stages, too few valid RT data points could be included due to high error rates to generate reliable RT measures.

RQ2: What are the roles of cognitive aptitudes (including working memory capacity, phonological STM, declarative memory ability, procedural memory ability, and musical aptitude) at different stages (time points) of learning oral Mandarin tonal word production, when the effect of ISI is controlled?

- RQ2-a: At the end of TS1:
    - Hypothesis 2a-i: WM capacity plays a facilitative role in oral picture naming accuracy.
    - Hypothesis 2a-ii: Declarative memory ability plays a facilitative role in oral picture naming accuracy.
    - Hypothesis 2a-iii: PSTM plays a facilitative role in oral picture naming accuracy.
    - Hypothesis 2a-iv: musical aptitude plays a facilitative role in oral picture naming accuracy.
- RQ2-b: At the end of TS3:
    - Hypothesis 2b-i: musical aptitude plays a facilitative role in oral picture naming accuracy.
    - Hypothesis 2b-ii: procedural memory ability plays a facilitative role in oral picture naming accuracy.

- Hypothesis 2b-iii: procedural memory ability plays a facilitative role in oral picture naming RT when controlling for L1 word naming RT.

RQ3: What are the effects of ISI and RI on the L2 *retention* of Mandarin tonal word production, when controlling for individual differences in cognitive aptitudes?

- RQ3-a: What are the effects of ISI and RI on the retention of the declarative component of Mandarin tonal word production, i.e., meaning-Pinyin associations?
    - Hypothesis 3a: the effects of temporal distribution of practice on the retention of declarative knowledge may be determined by the optimal ISI/RI ratios (10%-30%); therefore, Group A (ISI-1day; RI-1week) and Group D (ISI-1week; RI-4week) with optimal ISI/RIs are expected to outperform the other two groups, i.e., Group B (ISI-1day; RI-4week) and Group C (ISI-1week; RI-1week) on written picture naming accuracy.
- RQ3-b: What are the effects of ISI and RI on the retention of oral tone production skill, from Pinyin to articulation, on practiced words (i.e., words practiced in three training sessions)?
    - Hypothesis 3b-i: the massed practice schedule, which is more likely to enable participants to move from the second to the third stage of skill acquisition, should work better than the distributed practice schedule. Therefore, Groups A, B are expected to outperform Groups C and D in oral word naming accuracy on old words.

- Hypothesis 3b-ii: The massed practice groups are expected to respond faster than the distributed practice groups in Oral Word Naming RT on old words on the retention test, as increased/deeper procedural knowledge is likely to translate into faster RTs, and procedural knowledge is less vulnerable to memory decay during the RI.

- RQ3-c: What are the effects of ISI and RI on the retention of oral tone production skill (from Pinyin to articulation) on new words (i.e., words never practiced before)?
  - Hypothesis 3c-i: No strong hypothesis. As orally naming new words from Pinyin involves fusing tonal production with new combinations of segments, declarative knowledge about how to pronounce the tones is likely to be involved in this fusing process. As word naming on new words involves the retention of both declarative knowledge and procedural knowledge in articulation (and declarative knowledge is expected to be better retained in Groups A and D than Groups B and C, while procedural knowledge is expected to be better retained in Groups A and B than Groups C and D), Group A might outperform the other three groups in oral word naming accuracy on new words, at least descriptively.
  - Hypothesis 3c-ii: No predictions are made about the effect of RI, ISI, or their interaction, on oral word naming RT on new words.

- RQ3-d: What are the effects of ISI and RI on the retention of oral Mandarin word production skill, from meaning to articulation (as measured by Oral Picture Naming)?
    - Hypothesis 3d-i: No strong hypothesis. Oral picture naming involves the retention of both the declarative component of word knowledge (i.e., picture-Pinyin mappings) and the procedural component (i.e., oral articulation). Due to the double constraints of retaining declarative knowledge and procedural knowledge, Group A may outperform the other three groups on the picture naming accuracy on the retention test, at least descriptively.
    - Hypothesis 3d-ii: No predictions are made about the effect of RI, ISI, or their interaction, on oral picture naming RT on the retention test.

RQ4: What are the roles of cognitive aptitudes on L2 *retention* of Mandarin tonal word production when controlling for ISI and RI?

- RQ4-a: What is the role of cognitive aptitudes in the retention of *declarative word knowledge* (i.e., picture-Pinyin mappings)?
    - Hypothesis 4a: Declarative memory ability plays a facilitative role in written picture naming accuracy.
- RQ4-b: What is the role of cognitive aptitudes in the retention of *tone production skill on words practiced* (as measured by the Oral Word Naming task on old words)?
    - Hypothesis 4b-i: Musical aptitude plays a facilitative role in oral word naming accuracy on old words.

- Hypothesis 4b-ii: Procedural memory ability plays a facilitative role in word naming RT on old words (i.e., the higher procedural memory ability, the faster word naming RT) when controlling for L1 word naming RT.

- RQ4-c: What is the role of cognitive aptitudes in the retention of *tone production skill on new words*, (as measured by the Oral Word Naming task on new words)?
  - Hypothesis 4c-i: WM (complex) plays a facilitative role in word naming accuracy on new words.
  - Hypothesis 4c-ii: Musical aptitude plays a facilitative role in word naming accuracy on new words.
  - Hypothesis 4c-iii: Procedural memory ability plays a facilitative role in word naming RT on new words (i.e., the higher procedural memory ability, the faster word naming RT) when controlling for L1 word naming RT.
- RQ4-d: What is the role of cognitive aptitudes in the retention of *oral word production skills* (as measured by the Oral Picture Naming task)?
  - Hypothesis 4d-i: Declarative memory ability plays a facilitative role in oral picture naming accuracy.
  - Hypothesis 4d-ii: WM capacity plays a facilitative role in oral picture naming accuracy.
  - Hypothesis 4d-iii: PSTM plays a facilitative role in oral picture naming accuracy.

- Hypothesis 4d-iv: Musical aptitude plays a facilitative role in oral picture naming accuracy.
- Hypothesis 4d-v: Procedural memory ability plays a facilitative role in oral pic naming RT (i.e., the higher procedural memory ability, the faster oral pic naming RT) when controlling for L1 word naming RT.

Chapter 5: Methodology

5.1 Design of the Study

This study involves two between-subjects factors (i.e., ISI and RI), a number of individual difference covariates (i.e., working memory capacity, phonological STM, declarative memory ability, procedural memory ability, and musical aptitude) as within-subjects factors, and several outcome measures in word production (i.e., oral picture naming, written picture naming and oral word naming) as dependent variables.

Table 1. *Ratios of intersession intervals (ISI) to retention interval (RI)*

Conditions/ Groups	ISI (days)	RI (days)	ISI/RI
Group A	1	7	14%
Group B	1	28	4%
Group C	7	7	100%
Group D	7	28	25%

Table 1 presents the experimental manipulation of the two independent variables, i.e., ISI and RI, with two levels for each, the combinations of which create four experimental conditions. Following the design of Suzuki and DeKeyser (2015, 2016), the two levels of ISIs (1- vs. 7-day) and RIs (7-day vs. 28-day) were determined based on Rohrer and Pashler's (2007) optimal range of ISI and RI; that is, the ISIs were approximately 10% to 30% of the RI for optimal retention. It is worth noting that the optimal ratio of ISI and RI determined by Rohrer and Pashler was mainly based on the literature on the learning of declarative knowledge (such as facts and paired associates). This optimal ISI/RI ratio was used in the present study to (a)

replicate earlier findings on declarative learning of foreign vocabulary, and (b) test whether the optimal ratio for declarative learning can be generalized to the learning and retention of procedural knowledge or skills in L2 tone word production.

This study sets the ISIs and RIs at two levels each, i.e., 1-day (massed) vs. 7-day (distributed) for ISI, and 7-day (short-term) vs. 28-day (long-term) for RI, such that the ISI/RI ratio falls within the range of 10% to 30% for only one of the groups for the short-term RI and one for the long-term RI. Specifically, the optimal ISI/RI ratio occurs for the massed practice group with the short-term RI (i.e., Group A with a ratio of 14%) and for the distributed practice group with the long-term RI (i.e., Group D with a ratio of 25%). These two groups are expected to outperform the other two groups (Groups B and C) in the retention test of declarative word knowledge, because their ISI/RI ratios fall within the optimal range identified by Rohrer and Pashler (2007), while the ratios of the other two groups do not.

Table 2 shows the research design of the study. Participants were randomly assigned to the four experimental conditions. The four groups went through the same tasks/content of training and the same number of training sessions, but differed in terms of training or testing schedules (i.e., different ISIs or RIs). Each participant came in five times for this study on five separate days. The session on Day 1 was devoted to completing aptitude tests, after participants were informed about the study and filled out a background questionnaire. The session on Day 2 started with explicit instruction on Mandarin tones and tone changes in disyllabic words, and some tone practice on monosyllables. This session was then devoted to disyllabic word production training, which constituted Training Session 1. On Day 3 and Day 4,

participants came in for two more training sessions on disyllabic word production, one on each day. A quiz in the form of oral picture naming was administered at the beginning and the end of each training session, except for the beginning of the first training session, for which an oral word naming task was used instead, to keep track of learning within training sessions and of consolidation or more likely degradation of knowledge between training session intervals. After a retention interval, participants came in for the retention test on Day 5. As the session on Day 1 was not long enough to complete all the aptitude tests, three aptitude tests were administered on the later days, one each on Day 3, Day 4, and Day 5 (see Table 4 for the specific training schedule and procedures).

Table 2. *Research Design*

<table>
<tr><td></td><td>Day 1</td><td></td><td>Day 2</td><td></td><td>Day 3</td><td></td><td>Day 4</td><td></td><td>Day 5</td></tr>
<tr><td>Group A: ISI-1d; RI-1w</td><td rowspan="4">Aptitude Tests</td><td rowspan="4">→1 to 10 days depending on scheduling</td><td rowspan="4">Explicit Instructions + Training Session 1</td><td>→1 day</td><td rowspan="4">Training Session 2</td><td>→1 day</td><td rowspan="4">Training Session 3</td><td>→7 days</td><td rowspan="4">Retention Test</td></tr>
<tr><td>Group B: ISI-1d; RI-4w</td><td>→1 day</td><td>→1 day</td><td>→28 days</td></tr>
<tr><td>Group C: ISI-7d; RI-1w</td><td>→7 days</td><td>→7 days</td><td>→7 days</td></tr>
<tr><td>Group D: ISI-7d; RI-4w</td><td>→7 days</td><td>→7 days</td><td>→28 days</td></tr>
</table>

5.2 Participants

Participants had to meet each of the following requirements to participate: (1) speak English as their native language; (2) NOT be bilingual[1]; (3) NOT have prior knowledge of a tonal language such as Mandarin, Cantonese, Thai, Vietnamese, Ewe,

[1] Bilingual was defined as being able to speak two languages fluently. Those who reported having experience studying a foreign language in a classroom setting but not being able to speak a second language fluently were included in the study.

Krio, Twi, or Kikuyu; (4) NOT have more than two years of individual instruction in any combinations of musical instruments (including voice); (5) be between 18 and 40 years old; and (6) NOT have hearing loss or speech impairment. Those who did not meet any of the above criteria were screened out through email exchanges, or at the beginning of the first day of meeting through a participant background questionnaire (see Appendix A). One interested participant who was 41 years old was allowed to participate in the study. All were paid for their participation and gave informed consent.

Ninety participants began the study; however, nine did not come back after the first session. Among the 81 participants who completed all five sessions, one was excluded due to failure to follow instructions. Thus, a total of 80 participants provided data for this study.

The 80 participants (55 females) were all native speakers of English from the University of Maryland community (72 undergraduate, 7 graduate, and 1 post-bachelor), aged 18–41 (median 20, mean 21). None were bilingual, or spoke two languages fluently: eight participants reported no experience learning a second language, and 72 reported some experience learning a foreign language in a classroom setting. None had any prior knowledge of Mandarin or any other tone languages. With regard to musical experience, 57 participants reported that they had not had formal musical training experience (private music lessons), among whom 15 reported having group lessons from school. Ten participants reported having private lessons of less than two years, and one identified himself as having formal musical training and further noted that he was self-taught in two instruments (Saxophone and

Guitar). All can be considered as nonmusicians, following Wong and Perrachione's (2007) definition.

The sample of 80 participants was used to calculate reliability of the aptitude measures. Due to the two missing values for Ospan (the reason will be explained in 5.7.1), the reliability of Ospan was based on 78 participants. For Principle Component Analysis (PCA) on the cognitive aptitude measures, again, the sample of 78 participants were used. This sample was then also used to generate *z* scores of the aptitude variables for the subsequent hypothesis-testing analyses.

For hypothesis testing regarding the effects of ISI and RI and/or the role of aptitudes on learning and retention outcomes, ten additional participants were further excluded for the following reasons: eight were excluded because they reported in the post-study questionnaire that they had practiced outside of the training sessions (all participants were asked to not to practice outside of the study at the end of the first training session); one was discarded due to prior exposure to Mandarin Chinese via participating in a tone perception training study one month before the present study; and one more was dismissed due to lack of motivation to learn[2]. Data from these ten participants were excluded because their learning outcome was likely to be affected by factors other than the experimental condition or the cognitive aptitudes this study attempted to control.

After excluding the ten individuals described above, 68 participants were included for hypothesis testing that concerns the outcomes. The four groups ended up

[2] This participant insisted on wearing sunglasses in the lab during training, and yawned frequently during training. This participant's learning outcome also turned out to be an outlier for the group she was in.

with 18 participants in Group A, 18 in Group B, 16 in Group C, and 16 in Group D, so that the number of participants in the four groups was still well balanced. Table 3 summarizes the participants' background information regarding age of testing and gender. Participants in the four groups were comparable in terms of mean age (F (3, 64) = 0.454, p = .716), and gender distribution (Pearson Chi-Square X^2 = 3.524, p = 0.318).

Table 3. *Participants' information of the four experimental groups*

		Group A ISI-1d; RI-1w (n=18)	Group B ISI-1d; RI-4w (n=18)	Group C ISI-1w; RI-1w (n=16)	Group D ISI-1w; RI-4w (n=16)
Age	Mean	21.94	21.44	21.00	20.44
	Median	21	20	20	20
	Range	[18, 30]	[18, 35]	[18, 41]	[18, 30]
Gender	Female	9	13	12	12
	Male	9	5	4	4

5.3 Procedure

The study took place during the fall semester of 2016 and the spring semester of 2017, on the College Park campus of the University of Maryland. Each individual participated in five sessions of an hour or less on five separate days according to their specific training and testing schedules. The participants were randomly assigned to one of the four experimental groups (see Section 5.1 for the design of the study, including the ISIs and RIs for different groups).

Table 4 presents the training or testing procedures for the five sessions/days. This table serves as the roadmap about which tasks/activities were conducted on

which day (or in which session), in which order and how much time each one took. The specific procedure of each task will be detailed below in sections 5.5, 5.6 and 5.7.

Day 1 started with informing participants about the study, including the content and timeline, getting their consent to participate and asking them to fill out the participant background questionnaire. Interested participants who were not eligible to participate were screened out at this stage. This session was then devoted to cognitive aptitude testing. The cognitive tests administered in the first session include Shapebuilder, Nonword Repetition, Pitch STM, and Ospan (see details in 5.7.1, 5.7.2, and 5.7.3). These tests were administered in the same order to all participants. Day 2 started with explicit instruction on Mandarin tones (see details in 5.5.1), and then participants went through the first disyllabic word production training session (see details in 5.5.2). Day 3 started with the second disyllabic word production training session (see details in 5.5.3), and then participants completed a musical ability test (see details in 5.7.6). On Day 4, participants first went through the third and the last disyllabic word training session (see details in 5.5.4) and then completed the procedural memory test, i.e., the SRT task (see details in 5.7.5).

Table 4. *Training/Testing Procedures*

Days	Phases	Tasks/Activities	Length (min)
Day 1	Start	Informed consent and background questionnaire	5
	Cognitive tests	Shapebuilder	10
		Nonword repetition	5
		Pitch STM	15-20
		Ospan	15-20
	Instruction	Explicit instruction on Mandarin tones	5
		Tone practice in single syllables	10

Day 2		EI on tonal changes	3
	Training Session 1	Pre-test: Oral Word Naming	5
		Phase 1: Presentation (4 rounds)	8
		Phase 2: Declarative a) Presentation of Picture-Pinyin mappings (2 rounds) b) Picture-Pinyin ID task (1 round + review of incorrect items) c) Written picture-naming task (1 round)	12-14
		Phase 3. Procedural a) Oral word naming (2 rounds) b) Oral picture naming (2 rounds)	8-10
		Post-Session Quiz: Oral Picture Naming	3-5
Day 3	Training Session 2	Pre-Session Quiz: Oral Picture Naming	5
		Phase 1: Presentation (2 rounds)	4
		Phase 2: Declarative a) Picture-Pinyin ID task (2 rounds) b) Written picture naming task (1 round)	10-12
		Phase 3. Procedural a) Review declarative knowledge about tones b) Oral word naming (4 rounds) c) Oral picture naming (4 rounds)	20-25
		Post-Session Quiz: Oral Picture Naming	5
	Cognitive test	Musical ability test	5
Day 4	Training Session 3	Pre-session Quiz: Oral Picture Naming	5
		Phase 1: Presentation (2 rounds)	4
		Phase 2: Declarative a) Picture-Pinyin ID task (1 round) b) Written picture-naming task (1 round)	8-10
		Phase 3. Procedural a) Review declarative knowledge about tones b) Oral word naming (4 rounds) c) Oral picture naming (4 rounds)	18-22
		Post-session Quiz: Oral Picture Naming	3-5
	Cognitive test	SRT	10-12
Day 5	Cognitive test	CVMT (practice & the acquisition task)	10
	Retention test	Oral Picture Naming; Written Picture Naming; Oral Word Naming;	30

	Break	Take a break if retention test finished early, to fill the 30-min gap between the first and second parts of CVMT	
	Cognitive test	CVMT (delayed recognition & visual discrimination)	3
	End	Post-study questionnaire	5

Day 5 or the last session started with the declarative memory test, i.e., CVMT (see details in 5.7.4). As CVMT involves a 30-minutes delay between the first two components and the latter two components, after participants completed the first two components of CVMT, they took the retention test (5.6.1, 5.6.2, 5.6.3), which lasted about 20 to 25 minutes. Participants were asked to take a break to fill the gap. When the 30-minute delay was up, participants completed the third components of CVMT. This session, and the whole study for each participant, ended with a post-study questionnaire.

The post-study questionnaire was administered in a face-to-face interview, with the experimenter taking notes. It was first explained to them that, now that they had completed all training and testing for the study, the reason I was asking the following questions was because I wanted to be able to better interpret their data or results. The first interview question targeted at their declarative knowledge about the tones. They were asked to describe the four Mandarin tones and the rule about tonal changes in two-syllable words. I asked this question because I wanted to know, if they mispronounced a tone in the final retention test, whether it was because they did not remember how to say the tone or because it simply did not come out right. To put it in a formal way, if an error occurs in oral production in the final test, their answer to this first interview question can help differentiate whether the oral production error

was due to decayed (fuzzy, inaccurate, or incomplete) knowledge about tones or due to behavior failure (i.e., failure in executing the declarative knowledge). The second interview question inquired whether participants practiced the Mandarin words outside of the training sessions, although they were all asked to NOT practice at the end of the first training session. Their answers to this second question helped me exclude those who did not abide to the rule.

Some further detail is necessary here about how well the ISI and RI schedules were implemented in data collection. Among the final sample of 68 participants whose data were used for hypothesis testing regarding the effects of ISI and RI, six participants had slight deviations from their original schedules due to emergencies. The other 62 participants all came in on the scheduled dates according to their group assignment. The slight deviations were: 1 participant in Group A who had a RI of 8 days instead of 7 days; 2 participants in Group B who had a RI of 29 days plus 1 who had a RI of 31 days instead of 28 days; 1 participant in Group D who had an RI of 31 days instead of 28 days, and finally 1 participant in Group D who had an ISI of 8 days between the 2nd and 3rd training sessions instead of 7 days. All other intervals strictly respected the scheduled ISIs and RIs. These deviations of RI or ISI were so small that they hardly changed their ISI/RI ratios, and their effects were considered negligible.

5.4 Materials

Only real words, i.e., sounds (including the combinations of segments and tones) that do occur in real language use, were used in the present study, for both training and testing. When selecting words, care was taken to make sure that segments of both the initials (which are consonants) and the finals (which can

comprise up to three components, a prenuclear glide, a nuclear vowel and a coda nasal, with the first and last optional) do not present much difficulty for English NSs to pronounce. Tables 5 and 6 present the pools of initials and finals (together with their phonetic symbols in IPA, i.e., the International Phonetic Alphabet), used in the study, i.e., in the target words, the examples used for introducing tones, the items for tone practice in monosyllables, and the items in the generalization test. Note that the few sounds in italics (i.e., *s* [s] and *n* [n] as initials, and *ei* [əi] as finals) only occur in the word items in the generalization test, and do not occur in target words or words used to introduce tones or for practice before the formal learning of the target words, while the rest (10 consonant initials and 15 finals) all occur in the target words.

As can be seen from Table 5, the selected initial consonants almost all have a counterpart in English, and therefore do not present much difficulty for English NSs to pronounce, except for *h* [x], which is a voiceless *velar* fricative. The inclusion of this sound is not problematic for the purpose of this study, however, because [x] sounds very similar to the voiceless *glottal* fricative [h] in English, and it is written as *h* in Pinyin, which makes it easy for English learners to assimilate [x] with the English [h]. Assimilating [x] in Chinese directly with [h] in English is good enough for the purpose of the present study, which focuses on the learning of tones, rather than segments. Chinese only has [x] and does not have [h] as a phonological contrast. Thus, even if [x] is perceived and pronounced as [h], it does not cause any confusion for both learners and Chinese listeners. Due to its commonality in Chinese words, *h* [x] was included in this study. The rhymes selected for the study (see Table 6) should not present much difficulty either for native English speakers to pronounce.

Table 5. *List of initials used in this study*

Place/Manner	Unaspirated stops/ voiceless	Aspirated stops/ voiceless	Fricatives/ voiceless	Voiced
Labial	**b**	**p**	**f**	**m**
	[p]	[p^h]	[f]	[m]
Alveolar	**d**	**t**	***s***	**l, *n***
	[t]	[t^h]	[s]	[l], [n]
Velar	**g**	**k**	**h**	
	[k]	[k^h]	[x]	

Note: the two consonants in italics, i.e., *s* and *n* only occur in the word items for the generalization test, and do not occur in the target words.

Table 6. *List of finals used in this study*

Final/rhyme category	Finals
Mono-vowel	**a, o, i (y), u (w)** [a], [o], [i] ([j]), [u] ([w])
Multi-vowel	**ai, ao, ou, ua, uo** [ai], [au], [əu], [ua], [uo]
Nasal	**an, ang, ong, in, ing, en, *ei*** [an], [aŋ], [uŋ], [in], [iəŋ], [ən], [əi]

Note: (a) the two in brackets, i.e., "y and "w" are glides that can be considered variants of "i" and "u" used preceding the nuclear vowel; (b) the one in italics, i.e., *ei* only occur in the word items for the generalization test, and do not occur in the target words.

Consonants that are known to be hard for English NSs to pronounce as initials, i.e., *j* [tɕ], *q* [tɕ^h], *x* [ɕ], *z* [ts], *c* [ts^h], *zh* [tʂ], *ch* [tʂ^h], *sh* [ʂ], *r* [ɻ] were avoided. Likewise, vowels that can be hard for English NSs to pronounce, i.e., *e*, when it is pronounced as [ɤ] (which occurs when it stands alone as a single vowel for the final of a word, such as in *dé* [tɤ] (virtue)), and *ü* [y], were avoided. Although *e* [ɤ] was avoided, the other finals involving this letter in diphthong and nasal finals, i.e., *en* [ən], and *ei* [əi], were kept for use in this study.

5.4.1 Target words

As the learning task focuses on the four Mandarin tones in disyllabic words, there are 16 possible combinations of tone sequences of the four tones at the two positions (4 x 4) for disyllabic words (i.e., 1-1, 1-2, 1-3, 1-4, 2-1, 2-2, 2-3, 2-4, 3-1, 3-2, 3-3, 3-4, 4-1, 4-2, 4-3, 4-4). In order to limit the number of words to be learned to a manageable amount (cf. Chang & Bowles, 2015, for a tone-word perception training study, which had 24 target words), a decision was made to have one word for all 16 combinations, plus four more words for the T3-T3 combination because this combination involves tone sandhi/change. This decision resulted in a selection of 20 words (see Appendix B).

The target words were selected based on the following criteria: (a) they are real disyllabic words, (b) they are concrete nouns that are easy to illustrate pictorially, (c) the segments of the words are not difficult for English NSs to pronounce (piloted with a naïve native English speaker), and (d) each syllable occurs only once, carrying one particular tone, among the 20 target words, and therefore these 20 target words consist of 40 unique syllables.

Table 7 presents the list of 20 target words, together with their characteristics, including their phonetic symbols in IPA, English translation, tones and syllable structures of the first and second syllables. The four tones are equally distributed in the word-initial and word-final positions of the 20 target words. Tone 3 occurs eight times in the word-initial and word-final positions, respectively, and the other three tones each occur four times at each position. The eight occurrences of Tone 3 at the word-initial position are followed by Tone 1 once, Tone 2 once, Tone 4 once, and

Tone 3 five times. Five among the 20 words are T3-T3 combinations (e.g., lǎohǔ). These five words involve Tone 3 Sandhi, i.e., the Tone 3 at the word-initial position of these five words should be pronounced as Tone 2, i.e., the rising tone, as it is followed by a Tone 3. The other 11 occurrences of Tone 3 in these 20 words, i.e., three times when followed by a non-T3 tone, and eight times at the word-final position, should be pronounced as just a low tone.

Table 7. *Characteristics of the 20 target words*

No.	Pinyin	Phonetic symbols[3]	English translation	Tone_ Pos1	Tone_ Pos2	SylStr_ Pos1	SylStr_ Pos2
1	yīnghuā	jəŋxʷa	cherry blossom	1	1	GVN	CD
2	wūpó	upʰʷo	witch	1	2	V	CV
3	bānmǎ	panma	zebra	1	3	CVN	CV
4	wāndòu	wantəu	pea	1	4	GVN	CD
5	máoyī	mau'i	sweater	2	1	CD	V
6	yínháng	inxaŋ	bank	2	2	VN	CVN
7	píngguǒ	pʰʲəŋkʷo	apple	2	3	CVN	CD
8	tóufà	tʰəufa	hair	2	4	CD	CV
9	kǔguā	kʰʷukʷa	bitter melon	3	1	CV	CD
10	kǒnglóng	kʰʷuŋlʷuŋ	dinosaur	3	2	CVN	CVN
11	fěnbǐ	fənpʲi	chalk	3	3	CVN	CV
12	lǐngdài	lʲəŋtai	tie	3	4	CVN	CD
13	dàngāo	tankau	cake	4	1	CVN	CD
14	dìtú	tʲitʰʷu	map	4	2	CV	CV
15	dàogǔ	taukʷu	paddy	4	3	CD	CV
16	pùbù	pʰʷupʷu	waterfall	4	4	CV	CV
17	lǎohǔ	lauxʷu	tiger	3	3	CD	CV
18	gǎnlǎn	kanlan	olive	3	3	CVN	CVN
19	lǐpǐn	lʲipʰʲin	gift	3	3	CV	CVN
20	gǎngkǒu	kaŋkʰəu	harbor	3	3	CVN	CD

Note: [x] is voiceless velar fricative. As for the descriptors for syllable structures, C is a consonant, V is a vowel, D is a diphthong, N is a nasal, and G is a glide (i.e., [j] & [w]).

[3] Duanmu (2007) was followed in transcribing the syllables of the words.

The selected target words are comprised of syllables that represent a wide range of syllable structures in Chinese. The syllable structures of the target words include CV (4, 8; i.e., 4 at the word-initial position and 8 at the word-final position), CD (4, 7), CVN (8, 4), GVN (2, 0), V (1, 1) and VN (1, 0), where C denotes a consonant, V a single vowel, D a diphthong, N a nasal, and G a glide (i.e., [j] & [w]). Finals that are comprised of more than two vowels (such as *iao, iang, iong*) are avoided in the study because they may present pronunciation difficulty for English learners.

The auditory stimuli of the target words were recorded by two NSs of Mandarin, a female and a male, in a quiet room. The NS talkers were presented with characters of the disyllabic/ bimorphemic words (such as 老虎 for the word *lǎohǔ*) and were asked to read aloud the words naturally at a comfortable volume and pace. Each of the target words were read twice consecutively, and each talker completed three rounds of recording so that they could have a bit of practice and adjust their volume and pace. The audio recordings from the last round of recording from each talker were selected for use as auditory stimuli for training.

In this study, only one talker/voice matched with the gender of the participant was used for training. That is, female participants were trained with the audio recordings from the female voice only, and male participants were trained with the audio recordings from the male voice only. This was done for two reasons. Using only one talker/voice with each participant was for reducing talker variability presented to learners, because high talker variability (multiple talkers) may impede, rather than enhance, L2 learning of tonal contrasts at the initial stage, especially for

those with relatively weak perceptual abilities (Antoniou & Wong, 2016; Chang & Bowles, 2015; Perrachione et al., 2011). As females generally speak at a higher pitch range than male speakers, using the voice matched with the gender of the participants, rather than the voice mismatched with the gender of participants, was for providing models that are easier for participants to imitate for oral production practice.

Pictures were used to illustrate the meaning of each word, together with English translation at the word presentation stage. Only pictures were used in the later stages of training and outcome tests. See Appendix B for the complete set of visual stimuli (Pinyin spelling, English translation, and a picture for each word).

5.4.2 Words used at pre-training stages

Word examples were used when introducing Mandarin tones and tone changes in disyllabic words during the Explicit Instruction (EI) phases. The monosyllable used during the first EI phase was *ba* with the four tones. The disyllabic word example used during the second EI phase when introducing tone changes was *nǐhǎo* to illustrate Tone 3 Sandhi. The words that were used to practice hearing and producing tones in monosyllables immediately after introducing tones were the quadruplets of the following four monosyllables, i.e., *mo*, *bai*, *tan*, and *hong*, none of which overlap with any syllables of the target words and of the word items for the generalization test. The monosyllabic words for tone practice were also audio-recorded by the same female and male speakers.

5.4.3 Stimuli for the generalization test

A set of 64 new disyllabic words were selected for the generalization test (Gen words, thereafter) that aims at testing participants' ability in producing tones in new disyllabic words after training. Real words, instead of nonce words, were used, for the ease of later scoring, because it would be a much more straightforward rating task when the target words are real words rather than nonce words for native listeners.

The set of 64 Gen words are presented in Appendix C. This set consists of 4 lists of words; each list consists of 16 disyllabic words that constitute all 16 possible tone combinations for disyllabic words (i.e., 4 tones for each of the two positions, word-initial and word-final). When selecting the Gen words, care was taken to make sure that each constituent syllable with a particular tone (e.g., *fēn*) of the Gen words is unique among them and none of the constituent syllables with particular tones in the Gen words overlap with those of the target words. These words were carefully selected such that List 1 consists of words with *both* syllables from the set of 40 syllables of the target words, List 2 consists of words with only the *word-initial* syllables from the set of syllables of the target words, List 3 consists of words with only the *word-final* syllables from the syllables of the target words, and List 4 consists of words with both syllables new, not present in the target words. The *old* syllables in the Gen words (i.e. the syllables in the Gen words that overlap with syllables in the target words) carry different tones compared to their tones in the target words.

The new syllables of the Gen words (i.e., the second syllables of the words in List 2, the first syllables of the words in List 3, and both syllables of the word in List

4) are comprised of the initials and finals/rhymes (excluding tones) that make up the target words, with three new sounds, i.e., *s* [s] and *n* [n] as initials, and *ei* [əi] as final (see Tables 5 and 6), none of which should present difficulty at the segmental level. The only new rhyme that is not present in the target words is *ei* [əi], which occurs in four syllables of four Gen words.

5.5 Training Sessions

The focus of the training study was L2 learning of oral disyllabic word production in Mandarin. Before starting to be engaged in learning the target disyllabic words, participants were (a) first introduced to Mandarin tones, (b) engaged in some tone practice in monosyllables, and then (c) given explicit instruction on tonal changes in disyllabic words. After these pre-training steps (noted as *Explicit Instructions* in Table 2 for research design), participants started to engage in disyllabic word learning/practice, which was the focus of the present study. The pre-training steps only happened at the beginning of Day 2, after which they moved on to the first disyllabic word production learning session. Participants went through two more disyllabic word production training sessions on Day 3 and Day 4, one on each day. Each disyllabic word training session started and ended with a quiz in order to keep track of learning within sessions and enhancement or degradation between sessions.

In the following, I will first describe the pre-training steps, and then lay out the procedures for each of the disyllabic word production training sessions.

5.5.1 Pre-training steps

Step 1. Explicit Instruction on Mandarin tones in general. To start learning about Mandarin Chinese, participants were first given an introduction to Mandarin Chinese, particularly, Mandarin tones. See Appendix D for the EI sheet. This first instruction gives a general introduction to Mandarin tones. The instruction highlights the role of tones in distinguishing word meaning in Chinese, introduces the symbols used to mark the tones in Pinyin (i.e., the tone marks), and gives a description of each of the four tonal patterns and how to produce them in monosyllables. Note that Tone 3 is introduced as a *low* tone [21], that is, it starts low, then dips to the bottom of one's pitch range. The rising tail, depicted in dashed curve in the visual diagram, is introduced as optional (and only occurs when Tone 3 is spoken in isolation or at utterance-final position for emphasis). It is emphasized that Tone 3 is most often just a low tone in other contexts.

The instruction was printed on two pieces of paper, and audio-recorded by me, the experimenter. Participants were asked to hear the instruction while reading the EI sheet.

Step 2. Tone practice in monosyllables. Immediately after explicit instruction on Mandarin tones, participants engaged in tone practice in monosyllables. The practice at this stage serves to familiarize participants with the tones, and focuses on tone identification and production at the lower sound level. The syllables used for practice at this stage were *mo, bai, tan,* and *hong,* which represent most of the possible syllable structure types in Mandarin, i.e., CV, CD, and CVN. These four syllables, each with the four tones, make 16 words.

For tone identification practice, participants hear audio recordings of the 16 monosyllabic words, from a preprogramed list (see Appendix E), one at a time, and were asked to point at the Pinyin with the correct tone for each sound they heard on a paper sheet with the 16 monosyllabic words in Pinyin on it (see Appendix F for the sheet). The experimenter provided feedback for each trial before moving on to the next trial, by saying "Yes, correct" or "No, the answer should be…" while pointing at the correct answer. The set of 16 words were repeated in three rounds in different orders. Female participants were presented audio-recordings from the female voice and male participants were presented audio-recordings from the male voice, in this task, and for all training tasks throughout the study, i.e., including tone practice in monosyllabic words in this pre-training step and all practice in the three disyllabic word training sessions.

For tone production practice, participants were presented with the monosyllabic words in Pinyin on the computer screen, one at a time, and were asked to pronounce the words. The presentation of the words followed a preprogrammed list (see Appendix E), and the set of 16 words were presented in three rounds. Feedback, including a verbal judgment from the experimenter and a model pronunciation from the audio recording were provided to each of participants' productions. The feedback to the accurate productions were "yes, correct" plus an audio playing of the model pronunciation. The feedback to the inaccurate production were "no" or "close" depending on how accurate the production was, an identification of the problem (such as too high, too low, etc.), plus an audio playing of the model pronunciation.

Participants were asked to imitate the model pronunciation immediately after hearing it.

Note that Tone 3 in monosyllables for practice at this stage was presented as low-dipping (i.e., [212], with the rising tail) in auditory examples for perception practice. When participants were asked to practice producing it, they were required to produce a low tone (i.e., [21]), with the rising tail optional.

Step 3. Explicit Instruction on tone changes in disyllabic words. After tone practice in monosyllables, participants were then given explicit instruction on tone changes in disyllabic words. See Appendix G for the EI sheet. It introduces which tones remain the same and which tone changes. A big part of it regards variants of Tone 3 in disyllabic words; that is, Tone 3 becomes a rising tone when it is followed by another Tone 3 and just being a low tone in any other contexts (in the present study). The EI sheet was printed on paper and audio-recorded by me. Questions from participants were taken to make sure they fully understood the instruction.

These pre-training steps were only provided at the beginning of Day 2 before the first session on disyllabic word learning. These pre-training steps were NOT repeated on the later days.

5.5.2 Disyllabic Word Training Session I

Having being familiarized with tones in monosyllables and having learned the rules about tonal changes in disyllabic words, participants moved on to Disyllabic Word Training Session I. This training session (TS) begins with a pretest in oral word naming (from Pinyin to articulation) to assess how much participants could already do on disyllabic words given the instruction and practice they had received in the pre-

training steps. Participants then went on to learn the 20 target disyllabic words. It was introduced to the participants that the learning objective for them was to learn to orally produce the target words accurately and fluently when given the meaning of the words. As learning to orally produce words from meaning entails the learning of meaning-to-sound mappings and sound articulation, and sound was encoded as Pinyin in the written format to scaffold the memorization process, the task of oral word production learning can thus be broken down into two components: (a) the learning of the declarative component of word knowledge, i.e., meaning-Pinyin mappings, and (b) the learning of the procedural component in oral production. In addition, as words (including meaning and sound) need to be first presented for any learning to occur, a three-phase step-by-step training program was developed to scaffold the learning process: Phase 1 presentation of the target words (including meaning, Pinyin and sound for each word), Phase 2 declarative learning of word knowledge (i.e., picture-Pinyin mappings), and Phase 3 procedural learning of oral articulation of the words (from Pinyin and from pictures). The training session ended with a post-session quiz to assess their exit performance level by the end of the first training session.

Pre-test. An Oral Word Naming task was used to assess individual differences from the start in tone production ability in disyllabic words. This task consisted of the 20 target words and was administered before the words were even introduced. In this quiz, participants were presented with the 20 disyllabic words in Pinyin, one at a time, in a fixed random order, and were asked to pronounce the words aloud. They were given ten seconds to respond for each word, but they were asked to respond as quickly and accurately as they could. They were also asked to

avoid coughs, false starts, or hesitations, as once the microphone was triggered, the stimuli would disappear. Four English words were used as practice items to familiarize them with the task format. This task, and all following outcome tasks were administered using DMDX (Forster & Forster, 2003). Their oral productions were audio recorded. Response time (RT), i.e., the time from the start of presentation of stimuli to the start of articulation was also recorded.

Phase 1. Presentation of the target words. After the pretest, participants started to learn the 20 disyllabic words. They entered the presentation phase first. During this phase, participants were presented the 20 target disyllabic words on the computer screen one at a time. The presentation procedure for each word was as follows: the picture that illustrates the meaning of a word, together with its English translation (presented in font size 16 immediately below the picture), first appears at the upper center of the screen; after a one-second time interval, the Pinyin for the word (in font size 54) and an audio icon appear under the picture and English translation at the center of the screen, and at the same time, the sound/audio-recording for the word is played, twice, with a 0.75-second interval in between. The duration of the audio-recording for each word is less than one second. From the appearance of the Pinyin, all visuals (including the picture, English translation, Pinyin, and audio icon) stay on the screen for an additional four seconds, before the computer screen moves on to present the next word automatically. Participants were asked to pay careful attention to the sound of each word, how the sound is matched to the Pinyin, and the mappings between the picture, the Pinyin and the sound. Participants were asked to

not orally produce at this stage; instead, they were asked to focus on listening, viewing, and remembering the mappings.

The 20 target words were repeated four times in two blocks in the first disyllabic word training session. Within each block, the 20 words were divided into five sets, four words for each set, according to the tone combinations of the words. In the first block, words with the same tone for the *second* syllables were grouped into a set. In the second block, words with the same tone for the *first* syllables were grouped into a set. That is, the first block consisted of the following sets: Set 1 consisting of words with the tone combinations 1-1, 2-1, 3-1, and 4-1, Set 2 consisting of words with the tone combinations 1-2, 2-2, 3-2, and 4-2, Set 3 consisting of words with the tone combinations 1-3, 2-3, 3-3, and 4-3, Set 4 consisting of words with the tone combinations 1-4, 2-4, 3-4, and 4-4, and the final Set 5 consisting of words with the tone combination of 3-3. The following sets were formed for the second block: Set 1 consisting of words with the tone combinations 1-1, 1-2, 1-3, and 1-4; Set 2 consisting of words with the tone combinations 2-1, 2-2, 2-3, and 2-4; Set 3 consisting of words with the tone combinations 3-1, 3-2, 3-3, and 3-4; Set 4 consisting of words with the tone combinations 4-1, 4-2, 4-3, and 4-4; and finally Set 5 consisting of the additional four words with the tone combination 3-3. This was done in order to raise participants' awareness of the combinations of the tones. The words were first grouped according to the tone of the *second* syllable, because the second syllable is usually slightly longer than the first syllable in disyllabic words and therefore the tone of the second syllable can be assumed to be more salient. The presentation order within each block went from Set 1, to Set 2, Set 3, Set 4, and

finally Set 5; each set was repeated in succession two times. Within each set consisting of words with the same tone for the second syllable, the order of the four words always went by the tones of the first syllables, from T1, to T2, T3, and T4. Likewise, for the sets consisting of words with the same tone for the first syllable, the order of the four words always went by the tones of the second syllables, from T1, to T2, T3, and T4. The repetition of each set (consisting of only four words) in succession was done to increase awareness of the tone combinations and to facilitate remembering the mappings. Presenting words within each set according to the tone order of the other syllable aimed again at raising awareness about the regularity of the tone combinations of the words, which hopefully could help with the learning of these words.

Phase 2. Declarative learning of word knowledge. After the presentation stage of the target words, participants moved on to focus on the learning of the declarative component of word knowledge, i.e., meaning-spelling mappings of the words. The order of learning, practice, and testing will always go from meaning to form because the training study focuses on production. Participants were first presented with the target words, in pictures and Pinyin only (without sound), for two rounds. They were then tested in two tasks: a Picture-Pinyin identification task and a Written Picture Naming task, with feedback given. A Picture-Pinyin identification task is quick to administer and is useful to help establish Picture-Pinyin mappings. However, a mapping identification task may not be enough to force participants to remember each component of a word (including segment sequences and tone sequences) because participants can perform rather well in an identification task even

if they only remember partial component(s) of the word. For example, for *tie* in Chinese *lǐngdài*, if a participant remembers that this word in Chinese starts with *ling*, even if s/he does not even remember the tone for this first syllable or anything of the second syllable, s/he would be able to identify the Pinyin for this word, because each syllable for the 20 disyllabic words is unique. Therefore, in order to better prepare participants for oral word production when given a picture (which is the goal of word learning in the study), a written picture naming task was administered after the picture-Pinyin identification task, to force participants to remember each component of a word, including not only the segments, which may be easier for them, but also the tones for each syllable of the word.

As for the procedures of the tasks, during the two rounds of vocab presentation at this stage, the picture for the word (without English translation) first appears; after one second the Pinyin appears, and both remain on the screen for four more seconds. No sound was played so that participants focused on remembering picture-Pinyin mappings. The order of presentation of the words for the two rounds was random. Before starting vocab presentation at this declarative learning phase, participants were told that after the two rounds of presentation, they would be tested in a picture-Pinyin identification task, in which they will be tested until they reach a criterion of 80% accuracy, plus a written picture naming task.

In the picture-Pinyin identification task, participants were presented pictures for the target words, one at a time, on cards by the experimenter, and were asked to identify the Pinyin for the picture by pointing on a sheet with all 20 target words in Pinyin on it (see Appendix H). The experimenter provided feedback each time a

choice was made by saying “Yes, correct.” or “No.”/ “No, the answer is…” when pointing at the correct Pinyin. Due to time constraints, only one round of testing was administered; in this round of testing, 76.5% of the participants (i.e., 52 out of 68) reached criterion performance (80% accuracy). The items participants responded incorrectly were recycled one more time.

Participants then moved on to complete the written picture-naming task. In this task, participants were shown the pictures for the words on the computer screen, one at a time, in a fixed random order, and were asked to write down the Pinyin (including tone marks) for each picture on an answer sheet. For each item, after they wrote down their answer, they clicked to see the correct answer in blue on the screen as feedback. They were asked to compare the correct answer with their own, and make corrections if their answer was incorrect, before moving on to the next item. They were given 6 minutes to complete this task.

Phase 3. Procedural learning of oral production. After completing the written picture-naming task, participants entered Phase 3. For procedural learning of orally producing the words, participants practiced orally producing the words first from Pinyin (Oral Word Naming) and then from pictures (Oral Picture Naming). In oral word naming, participants were shown the target words in Pinyin on the computer screen, one at a time, and asked to pronounce the words. In oral picture naming, the pictures were presented instead, one at a time, and participants were asked to produce the words. The 20 target words were practiced twice in oral word naming and twice in oral picture naming. For oral word naming practice, the order of presenting words followed the order in the first block in Phase 1; that is, the words

were presented in sets (as described in Phase 1), by the order of the tones of the second syllable of the words, with each set repeated in succession before moving on to the next set. For oral picture naming practice, the order of presenting pictures followed the order in the second block in Phase 1; that is, the words were presented again in sets, by the order of the tones of the first syllable of the words, with each set repeated in succession before moving on to the next set. Again, this systematic ordering was done to facilitate learning.

Feedback, again, including a verbal judgment from the experimenter and a model pronunciation from the audio recording was provided to each of participants' productions. The feedback to accurate productions was "yes, correct" plus an audio-recording of the model pronunciation. The feedback to the inaccurate productions was "no" or "close" depending on how accurate the production was, an identification of the problem (such as too high, too low, etc.), plus an audio playing of the model pronunciation. Participants were asked to imitate the model pronunciation immediately after hearing it.

Post-session quiz. The training session ended with a post-session quiz in oral picture naming (as this was the target skill) to measure their exit performance in oral production from meaning to articulation. The task procedure was the same as the oral word naming task except that (a) pictures for the target words were presented instead of Pinyin; (b) pictures for the four English words were used for practice items; and (c) the test items were presented in a different fixed random order.

5.5.3 Disyllabic Word Training Session II

Pre-session quiz. The second disyllabic word training session started with a pre-session quiz in oral picture naming. The procedure of this quiz was the same as the oral picture-naming quiz administered at the end of the first training session except that the test items were presented in a different fixed random order.

Phase 1. Presentation of the target words. This phase served as a review of the words learned in the first disyllabic word training session. The procedure for presenting the words was the same as it was in Phase 1 in the first training session, except that the target words were presented only twice, in two rounds, first by the order of the tones of the 2nd syllable in Round 1 (i.e., 1-1, 2-1, 3-1, 4-1, 1-2, 2-2, 3-2, 4-2, 1-3, 2-3, 3-3, 4-3, 1-4, 2-4, 3-4, 4-4, 3-3, 3-3, 3-3, 3-3), and then by the order of the tones of the 1st syllable in Round 2 (i.e., 1-1, 1-2, 1-3, 1-4, 2-1, 2-2, 2-3, 2-4, 3-1, 3-2, 3-3, 3-4, 4-1, 4-2, 4-3, 4-4, 3-3, 3-3, 3-3, 3-3).

Phase 2. Declarative learning of word knowledge. In this second training session, Phase 2 started directly with the picture-Pinyin ID task, without the two rounds of presentation of the picture-Pinyin mappings as in Phase 1 in the first training session. The presentation step was removed because it was not necessary after the review of words in Phase 1 in this training session. The procedure of the picture-Pinyin ID task was the same as it was in the first training session except that all the target words were tested in two rounds, in random orders. In the second round of picture-Pinyin ID testing, 88.2% of the participants (i.e., 60/68) achieved 100% accuracy; seven participants missed only one out the 20 words, and only one participant missed two words. After the picture-Pinyin ID task, participants

completed the written picture naming task, with the same procedure as it was in the first training session expect for a different ordering of the items.

Phase 3. Procedural learning of oral production. Phase 3 in this training session started with a review of declarative knowledge about the tones. Participants were first asked to verbally describe each of the four Mandarin tones and how they act or change in two-syllable words. Feedback was given and corrections were given when necessary. After the review step about tones, participants then focused on oral production practice in this session. They engaged in 4 rounds of oral word naming practice and then 4 rounds of oral picture naming practice, the amount of practice doubled as compared to the amount in the first training session. The procedures for the practice activities in this training session were the same as they were in Phase 3 in the first training session. As for the orders in presenting the words for practice, the first two rounds were the same as they were in TS1 for both oral word naming and oral picture naming; for the latter two rounds, the words were ordered by tone pair contrasts (e.g., 1-2 & 2-1, 2-2 & 4-4, 3-2 & 2-3) scattered or separated by other tone pairs (e.g., 3-3, 1-1). Items in Round 3 for both oral word naming and oral picture naming practice were ordered according to the following tone pair contrast pattern: 1-2, 2-1, 3-3, 2-2, 4-4, 3-3, 3-2, 2-3, 1-1, 4-2, 2-4, 3-3, 3-1, 1-3, 3-3, 4-1, 1-4, 3-3, 3-4, 4-3. Items in Round 4, for both oral word naming and oral picture naming, were presented in an order that reversed the order of the tone pairs, i.e., 2-1, 1-2, 3-3, 4-4, 2-2, 3-3, 2-3, 3-2, 1-1, 2-4, 4-2, 3-3, 1-3, 3-1, 3-3, 1-4, 4-1, 3-3, 4-3, 3-4.

Post-session quiz. An oral picture-naming task was administered as a post-session quiz. Again, the procedure was the same as for the oral picture-naming quiz

administered at the end of the first training session and the beginning of this session, except that testing items were presented in a different fixed random order.

5.5.4 Disyllabic Word Training Session III

Pre-session quiz. An oral picture-naming task was administered as the pre-session quiz, with the same procedure as the previous oral picture naming tasks, except for a different presentation order of the testing items.

Phase 1. Presentation of the target words. This phase, again, served as a review of the target words. This phase in this session was almost the same as Phase 1 in TS2, except the order of presenting words. In this phase of this training session, Round 1 was the same as Round 2 in TS2-Phase1, and Round 2 the same as Round 1 in TS2-Phase1.

Phase 2. Declarative learning of the target words. This phase in this training session was almost identical with Phase 2 in TS2, except that in the picture-Pinyin ID task, the set of target words were only tested in one round (instead of two rounds as in TS2). In this one round of identification testing, 97.1% of the participants (i.e., 66/68) achieved 100% accuracy; only one participant missed one word and another participant missed two words. After the picture-Pinyin ID task, participants completed the written picture-naming task, with the same procedure as it was in the previous training sessions expect for a different order in presenting the items.

Phase 3. Procedural learning of oral production. This phase in this training session again started with a review of declarative knowledge about the tones, as in the same phase in TS2. Participants then focused on oral production practice. As in TS2,

participants engaged in 4 rounds of practice in oral word naming and then 4 rounds in oral picture naming. For both oral word naming and oral picture naming practice, items in Round 1 were presented in sets according to the tone of the 2nd syllable of the words (1-1, 2-1, 3-1, 4-1, 1-2, 2-2, etc.), and items in Round 2 were presented in sets by the tone of the 1st syllable of the words (i.e., 1-1, 1-2, 1-3, 1-4, 2-1, 2-2, etc.). In Round 3 and Round 4, for both oral word naming and oral picture naming practice, items were presented in a randomized order. The remainig procedures were the same.

Post-session quiz. The third training session again ended with a post-session quiz in oral picture naming, with the same procedure as the previous oral picture naming tasks, except for another different presentation order of the testing items.

5.6 Final Outcome Tests

Three tasks, i.e., an oral picture-naming task, a written picture-naming task, and an oral word-naming task, were administered as retention tests after an RI. The picture naming tasks, either oral or written, focuses on the 20 target words. As for the oral word-naming task, in addition to the 20 target words, a set of *new* disyllabic words were included, to test participants' ability in generalizing their tonal ability in reading new words. The generalization test in oral word naming was used to check whether systematic learning of the tones, which cannot be due to just item learning, occurred, and whether the effects of practice distribution generalized to new words. The task order always went from the higher word level to the lower sound level, i.e., from oral picture naming, to written picture naming, and then oral word naming, in order to minimize the influence of practice from the lower level to performance at the higher level. The oral picture-naming task and the oral word-naming task were

computer-administered by using DMDX software, with both response (oral production) and response time recorded.

5.6.1 Oral Picture Naming

The oral picture-naming task was designed to tap into word production from conceptualization to articulation, which was the target skill for training. In this task, participants were presented with the pictures of the target words, one at a time, and were asked to orally name the pictures in Mandarin. To correctly produce the word when seeing a picture, participants had to (a) retrieve the correct word (including the segments and tone sequence) from memory, and (b) accurately articulate it. This test consisted of 20 items.

In this oral picture naming task, each trial starts with a fixation "*" in the center of the screen for 500 ms, which was then replaced by a picture. The picture stays on the screen until the participant starts saying the word into the microphone. The picture disappears as soon as the microphone is triggered by the start of the utterance. After a 2.5-second blank screen, the fixation asterisk appears again, signaling the beginning of the next trial. Participants were given 20 seconds to respond for each item, but were asked to respond as quickly and accurately as possible. Oral productions were audio-recorded. RT from the onset of the presentation of the visual stimuli till the point when the microphone was triggered by an oral response (or a sound in general) was recorded. Participants were presented with four practice items using English examples at the beginning to familiarize them with the task format. Before the start of the practice items, the researcher worked with the participant to adjust the sensitivity of the threshold for the input signal strength

and the threshold for the sibilant signal strength so that it did not respond to the ambient noise or non-speech sounds but did respond to the perceived onset of the participant's production for the stimuli. In addition, participants were instructed to avoid coughs, false starts, or hesitations, and speak clearly into the microphone. During this test, the experimenter was standing at the back of the room, observing, and taking notes if any false triggering happened; the RTs from false triggers are not accurate and thus have to be excluded for later analysis. The test items were presented in a fixed order, which was done to make it easier and less error-prone for the experimenter to keep track of items with false triggers as the test progressed.

The procedures for the five oral picture naming tasks conducted at the beginning or the end of the training sessions were the same as this oral picture naming task except that participants were given 10 seconds to respond in the previous five tasks, but 20 seconds in this one (to allow for enough time for lexical retrieval after the retention interval).

5.6.2 Written Picture Naming

The written picture naming task was designed to test the declarative component of word knowledge, i.e., meaning-spelling mappings of the words. Participants were presented the pictures for the target words, one at a time, and were asked to write down the Pinyin (including tone marks) for each picture on an answer sheet. Each trial started with a ding sound, which was used to gather their attention to view the screen. After the ding sound, a picture, with its item number directly above it, appeared at the center of the screen. Participants were given a maximum of 30 seconds for each item. They were instructed that when the 30-second limit was

reached, the screen would move on to present the next item automatically, and if they finished early, they should press the spacebar to continue. The test items were presented in a fixed random order.

5.6.3 Oral Word Naming

The oral word-naming task was designed to test participants' word production ability at the lower sound level from word form to articulation, without the involvement of word meaning. In this task, participants were presented with the words in Pinyin, one at a time, and asked to read them aloud. There were two subtests or blocks in this task. The first block consisted of the 20 learned target words. The second block was the generalization test, consisting of 64 new disyllabic words in four lists (see the materials section for the description; see Appendix C for the stimuli). The 64 new words were presented in four sub-blocks, one list of 16 words in each sub-block, with the order from List 1, to List 2, List 3, and List 4. Participants were given brief breaks between the sub-blocks. The task procedure was the same as for the oral picture naming task, except that words in Pinyin were used as visual stimuli (instead of pictures) and participants were given 10 seconds to respond in this task (instead of 20 seconds in oral picture naming).

L1 Word Naming. As L2 Mandarin word production RTs (including the RTs in oral picture naming and oral word naming) will be used as outcome measures to investigate the effects of ISI and RI (between-subject variables), and individual participants can simply vary in word production response time (some are slower and some faster) even in their L1, an L1 word naming task, using 30 high-frequency disyllabic English nouns that are matched with the target Mandarin words in length

(see Appendix I), was used to control this individual variation in basic cognitive processing speed. This L1 word naming task administered before the Mandarin oral word naming task described above. The RTs from the L1 word naming task will be used as a covariate to account for individual differences in L1 word production RT when L2 word production RTs are used as dependent variables.

5.6.4 Scoring Procedures for the outcome tests

Oral Picture/Word Naming tasks. The oral Mandarin word productions, from the oral picture naming tasks (administered as part of the retention test on Day 5 or as quizzes at the beginning or the end of the training sessions) and from the oral word naming tasks (in the pretest or in the retention test, old words and new words), were all scored according to the same criteria. Answer sheets containing the target words in Pinyin in an order corresponding to the presentation order for each of the tasks were used for scoring. Each oral production of a two-syllable word received a full score only if each component of the word was correctly pronounced. In other words, the evaluation of each word (disyllabic) was based on the evaluation of correctness of each of the constituent components. Each two-syllable word was broken down into four components for evaluation, i.e., 1st syllable segments, 1st syllable tone, 2nd syllable segments, and 2nd syllable tone, and each component was allocated 1 point, so the total maximum score for each trial was 4 points. *Tone* was scored dichotomously, as either correct (√, 1 point), or incorrect (×, 0), with the criterion for correct being sounding clearly native-like. If the tone sounded strange, even only a little bit, it was scored as incorrect, and no partial points were given. The segments for each syllable were allocated 1 point, and each syllable segment (e.g.,

ping, the first syllable for the word *pingguo*) was further broken down into two subcomponents, i.e., the initial (*p*) and the final (*ing*), with 0.5 point allocated to each subcomponent. That is, when both the initial and the final subcomponent for a syllable was pronounced correctly, the segment component of the syllable was scored as *correct* (√) and rewarded 1 point. When only the initial or only the final was pronounced correctly, the segments component of the syllable was scored as *half correct* (√`) and given only 0.5 point. When neither the initial nor the final was pronounced correctly, the segments component to the syllable was scored as *incorrect* (×) and given 0 point.

When no response to a given item was observed, two situations were differentiated: if no response was due to a false trigger (which had been noted down by the experimenter during testing), so that participants did not get a chance to say the word before the computer moved on to the next item, those nonresponses were treated as missing data; in the other situation in which no response was due to the fact that they did not remember or did not know how to say the word, those nonresponses were scored as incorrect (0). Where there were false starts or self-corrections, their first try was scored. See Appendix J for the specific scoring rubrics for scoring tones and segments of oral productions (i.e., subsection D. Tone Scoring Rubric and subsection E. Segments Scoring Rubric under Appendix J).

The researcher, who is a native speaker of Mandarin Chinese, scored all audio-recordings of Mandarin word productions from all participants. A second rater, who is also a NS of Mandarin Chinese, scored 15% of the data (i.e., 10 out of the 68 participants). The second rater was first trained by the researcher on how to score (see

Appendix J for the instruction sheets provided to the second rater), practiced by scoring some samples during training (36 trials) with feedback provided, and then scored audio-recordings from 10 participants independently. The inter-rater reliability in terms of Pearson correlation was .920 for the tone composite score (i.e., the sum of the two tone component scores for each trial), .925 for the segment composite score (i.e., the sum of the two segment component scores for each trial), and .939 for the 4-component composite score, which indicated excellent interrater reliability; Thus, only the scores from the researcher who finished scoring all data were used for further analyses.

Written Picture Naming task. The answer sheets from the written picture naming task administered as part of the final retention test was scored by the researcher. Again, each trial, or two-syllable word, was allocated 4 points in total, 1 point for each of the four components, i.e., 1st syllable tone mark, 1st syllable segments spelling, 2nd syllable tone mark, and 2nd syllable segment spelling. Again, the tone marks were scored dichotomously as either correct or incorrect; the segments spellings were further broken down into initials and finals, and half points were given when only the initials or the finals were put correctly.

When scoring was finished, the component scores (1, 0.5, or 0) for each trial in each task (all oral picture naming tasks, oral word naming tasks, and the written picture naming task) for all participants were entered to excel sheets (See Appendix J, subsection C for an example of the structure and format for data entering) for calculating task scores for each participant. Depending on task type, a composite score of different components was calculated as the trial score for each task. For the

oral word naming tasks (i.e., the pretest on Day 2 and the oral word naming tasks on old and new words on Day 5), a composite score of the two *tone* components (one for each syllable of a disyllabic word) was calculated as the trial score, because this task type was used to asses tone production skills. For the oral picture naming tasks (i.e., the five ones either at the beginning or the end of each training sessions, and the one on Day 5), and the written picture naming task (on Day 5), a composite score of the *four* components was summed for each trial in these tasks, because these task types focused on not only the tone components but also the segmental components. An accuracy rate was then calculated, for each participant on each task, as that participant's task score.

5.7 Aptitude Tests

5.7.1 Working Memory Tests

Shapebuilder. The web-administered Shapebuilder task (Atkins et al., 2014) was used as a measure of visuospatial working memory. In this test, participants were asked to remember the order and spatial position in which a series of colored geometric shapes were presented. Participants saw a four-by-four grid of connected squares with four shapes (squares, circles, triangles, and diamonds) in four colors (yellow, green, red, and blue) lining each of the four sides of the grid (see Figure 4). For each trial, participants were presented a sequence of colored shapes (between 2 to 4) that appeared one at a time in one of the 16 grid positions. After the last shape was presented, participants were asked to recreate the sequence by clicking on the correct colored shape and dragging it to the appropriate spatial location. There were 26 trials,

with 6 having 2 stimuli per trial, 9 having 3 stimuli per trial, and 11 having 4 stimuli per trial. The test started from trials with the shortest length, i.e., 2 stimuli per trial, and continued with trials with longer lengths, i.e., 3 stimuli per trial, and then 4. Within each set of trials of a given trial length, the trial difficulty increased by increasing the variation of colors and/or shapes of the stimuli. Participants received immediate feedback about the accuracy of each item by viewing the points awarded to each item immediately after the participant released the mouse button. Participants were informed ahead in the instructions that the number of points they earned would increase the more they got correct without making a mistake.

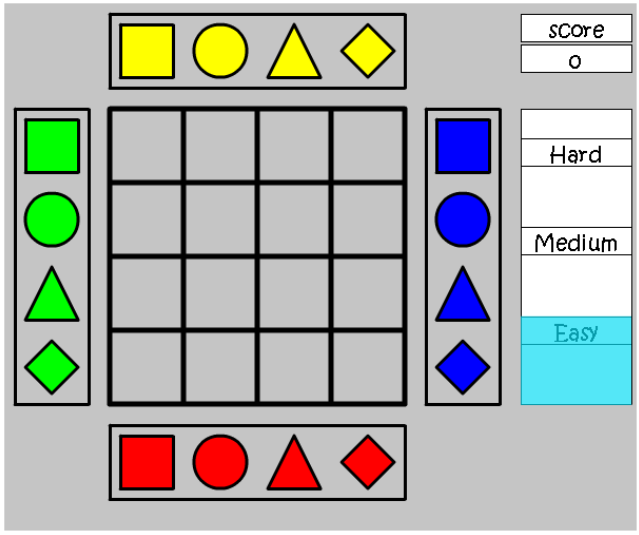


Figure 4. Screen shot of the Shapebuilder task

The scoring procedure was as follows: 15 points were rewarded to an item if the correct shape with the correct color was placed into the correct location, 10 points for any item with the correct shape placed in the correct location, but with incorrect color, and 5 points for only getting the location correct, and 0 if the location was wrong. In addition, the points rewarded to the items within a trial went up

exponentially, i.e., 15 points for getting the first item of a trial correct, an additional 30 points for getting the second item correct after getting the first item correct, an additional 60 points for getting the third item correct after getting the first two items correct, and an additional 120 points for getting the fourth items correct after getting the first three items correct. The maximum score for this test was 3,690 points.

Operation Span. The automated operation span task (Ospan) from Unsworth, Heitz, Schrock, and Engle (2005) was used as a test of verbal working memory. In the Ospan task, participants first saw a math problem (e.g. (1*2) + 1 = ?) and were asked to solve it as accurately and quickly as they could. They were instructed to click the mouse once they knew the answer to advance to the next screen, and saw a number (e.g., "3") and were required to judge if the number was the correct solution by clicking on "True" or "False". After each math problem, they were presented with an English letter for 800 ms, and were asked to remember it. After each set of math problems and letters (which can vary from 3 to 7 in set size), they were required to recall the letters from the presented order by clicking on the correct box next to the appropriate letters in the correct order. There were 15 trials in total, three trials for each of the five set sizes (i.e., 3-7). The total number of letters to be recalled was 75. Participants were instructed to keep their math accuracy at or above 85% all the time in order for their results to be valid.

At the end of the task, the program automatically generated two span scores: the absolute span score, which is the sum of the letters in all perfectly recalled sets, and the total number correct, which is the total number of the letters recalled in the correct position. To illustrate, if an individual correctly recalled 3 letters in a set size

of 3, 4 letters in a set size of 4, and 4 letters in a set size of 5, their absolute Ospan score would be 7 (3+4+0), and the total number correct would be 11 (3+4+4). The absolute Ospan score was adopted in the study. In addition, an 85% accuracy criterion (i.e., a maximum of 12 errors out of the 75 operations) was applied to all participants, following the original study that validated this automated Ospan task (Unsworth et al., 2005), in order to make sure the task was performed appropriately. Scores from two participants were discarded due to this criterion. The total number of participants remaining for this task was 78.

5.7.2 Phonological Short-Term Memory Test

Nonword repetition. Nonword repetition (NWR) was used as a test of productive PSTM. The test was adopted from Martin and Ellis (2012) who borrowed the stimuli from Gathercole, Pickering, Hall, and Peaker (2001). See Appendix K for the complete set of stimuli used in this test. For each trial, participants heard a list of one-syllable consonant-vowel-consonant nonwords based on English phonotactics and were asked to repeat them as accurately as possible. There were altogether 16 trials or lists of nonwords, i.e., four lists at each of the four lengths: three, four, five, and six nonwords. Participants heard the lists in the same order, beginning from the shortest lists and continuing with lists of increasing length. Participants' responses to all items were audio-recorded. Following Martin and Ellis (2012), scoring was done on a phoneme-by-phoneme basis by a trained NS of English. A second rater who is also a NS of English scored 25% of the samples (20 out of 80), and the inter-rater reliability in term of Pearson correlation was .879. The maximum number of phonemes recalled correctly on any repetition list was calculated for each participant

and was taken as their NWR score. The maximum possible score was 22 correct phonemes.

5.7.3 Pitch STM Test

Pitch STM. The pitch STM test adopted from Bowles et al. (2016) was used as a measure of pitch STM. The auditory stimuli used in this test were non-speech level tones, i.e., sine waves. Pitch STM was tested in two conditions. In the first condition, after a warning sound indicating the beginning of each trial, participants heard the first tone, after a 5-second silence the second tone, and then were required to press a button to indicate whether the second tone was the same or different from the first. In the following second condition, i.e. the interference condition, trials were the same in structure except that six intervening tones were played during the 5-second interval between the two tones. There were 48 items in each of the two conditions. Percent accuracy in each of the two conditions was calculated (i.e., PitchSTM-control and PitchSTM-interference).

5.7.4 Declarative Memory Test

Continuous Visual Memory Task. Following Morgan-Short et al. (2014) and Carpenter (2008), the Continuous Visual Memory Task (CVMT, Trahan & Larrabee, 1988), which was designed to minimize reliance on verbal encoding strategies, were used to test visual declarative memory ability. This test has four components: practice, an acquisition task, a delayed recognition task, and a visual discrimination task. In the acquisition task, participants viewed a series of complex abstract designs one at a time, and were then assessed on their ability to recognize

seven of the designs (i.e., the target items) from a mixed list of "old" and "new" designs. The "old" items, i.e., the seven target designs, were presented seven times (which totals 49 trials) interspersed among 63 "new" distractor items that appeared only once. All items (N_{item}=112) were presented in a fixed randomized order to all participants. Participants were asked to state aloud whether the item is "old" or "new". The experimenter wrote down participants' responses on answer sheets. After the acquisition task was completed, there was a 30-min delay, and then participants completed a delayed recognition task (N_{item}=7). The final task was a visual discrimination task (N_{item}=7) that was used to distinguish visual discrimination deficits from visual memory ability. All participants scored perfectly (7 out of 7) on the discrimination task, suggesting that none had visual discrimination deficits. A *d*-prime score for the acquisition task was calculated for each participant as a CVMT acquisition score. In addition, the number of correct responses in the delayed recognition task was calculated as a CVMT delayed score.

5.7.5 Procedural Memory Test

Serial Reaction Time Task. The serial reaction time (SRT) task was used to measure implicit sequence learning in procedural memory, following Lum, Conti-Ramsden, Page, and Ullman (2012), Conti-Ramsden, Ullman, and Lum (2015), Hamrick (2015), and Tagarelli, Ruiz, Vega, & Rebuschat (2016). The probabilistic version of the SRT task adapted from (Kaufman et al., 2010) was used in this study. In this task, four horizontally arranged white squares were shown at the center of the computer screen. For each trial, participants saw a black dot appear in one of the four squares, and were asked to respond by pressing the key corresponding to the location

of the black dot as quickly and accurately as possible. After the correct key was hit, the black dot moved on to another location. Unknown to participants, the sequence of stimuli followed a probabilistic rule. In each task block, training trials were interspersed with control trials, with the former generated by Sequence A (1–2–1–4–3–2–4–1–3–4–2–3), which occurred with a probability of 0.85, and latter generated by Sequence B (3–2–3–4–1–2–4–3–1–4–2–1), which occurred with a probability of 0.15. In both sequences, each location (i.e., 1, 2, 3, 4) and first-order transitions (e.g., 1-2, 1-3, 1-4) occurred with the same likelihood. However, the two sequences differed in the second-order conditional information that led to a different prediction (or successor) in each sequence. For instance, the two-digit sequence 1-2 was always followed by 1 in Sequence A, but it was always followed by 4 in Sequence B. There were eight blocks of trials, 120 trials for each block, which totaled 960 trials.

The SRT task was scored according to the method in Kaufman et al. (2010). First, error responses (2.5% of trials) were discarded, as well as outliers more than three standard deviation from the mean that was calculated individually for each block and participant (1.6%). Then, the average learning effect, i.e., the global effect size (Cohen's *d*) based on the whole sample of participants across blocks 3-8 (blocks 1 and 2 were not included because learning was not established in the first two blocks) was calculated, by comparing the mean RTs for probable vs. improbable trials across those blocks. Cohen's *d* was .26 in this sample (.19 in Kaufman et al., 2010). The next step was to assess whether participants showed a learning effect as least as large as the average learning effect (calculated above) in each block (blocks 3-8) for each participant. For each block (blocks 3-8), if a participant's mean RT for probable

trials was less than the difference between his/her mean RT for improbable trials and .26 times his/her standard deviation of RT on improbable trials, then he/she received a score of 1 for that block. If not, he/she received a score of 0 for that block. The total score for each participant was calculated by summing up their scores across blocks 3-8. This total score, ranging from 0 to 6, was taken as the SRT score. The split-half reliability (using Spearman-Brown correction) of the SRT scores from this sample was .666 (.44 in Kaufman et al, 2010), and the distribution was normal.

5.7.6 Musical Ability Test

The productive tonal memory test developed by Slevc and Miyake (2006) was used to measure productive tonal musical ability. In this test, a musical tune (2-7 notes long) was played twice for each trial, and participants was asked to reproduce the tune from immediate memory. Participants were asked to sing the tunes as accurately as possible in terms of pitch. They were asked to sing, rather than hum, and to use the syllable "*la la*….." in order to avoid different behaviors that may be hard to compare or analyze. There were 18 items in this test, 3 items for each tune length. Participants' singing was audio-recorded.

The audio-recordings were hand-scored by using the Praat computer program (Boersma & Weenink, 2015). The Praat program generates the frequencies (Hz) of the recorded singing. Slevc and Miyake's (2006) scoring criterion was followed. A note was considered accurate if the frequency of the stable part of the pitch was within one semitone of the target. The frequencies of the notes from different octaves were used as appropriate for female and male voices. Each item was scored dichotomously. For example, an item consisting of a 5-note tune was scored as

correct only if all five notes produced for this tune were accurate; it is scored as incorrect if one or more notes were produced inaccurately. The researcher scored 46 of the 80 samples, and a trained research assistant scored 34 samples. The accuracy rate for each participant was calculated.

Chapter 6: Results

This chapter presents the results of this study. Descriptive results on the set of cognitive aptitude measures and on the learning outcome measures are presented first. Next, the results in response to the four research questions regarding (a) the effects of ISI on L2 learning across training sessions, (b) the role of aptitudes at different stages of L2 learning across training sessions, (c) the effects of ISI and RI on retention performance, and (d) the role of aptitudes on retention performance, are reported respectively, in sequence.

6.1 Cognitive Aptitudes

6.1.1 Descriptive statistics of the aptitudes measures

Table 8 presents the descriptive statistics of the nine aptitude measures, including the number of participants, the possible maximum score, mean, standard deviation, minimum and maximum scores and the reliability of each measure. Table 9 presents the distribution of each of the aptitude measures, including the Skewness and Kurtosis statistics, standard errors for each, and the calculated *z* scores. The *z* scores were used as criterion to determine distribution normality of each of the measures – if the *z* score, for either skew or kurtosis, was greater than an absolute value of 1.96 ($p < .05$), the distribution was considered non-normal and transformations were then applied.

Shapebuilder. The possible maximum score for Shapebuilder was 3690 points. The mean was 1625.75 (*SD* = 415.4) from the sample of 80 participants in this study, with the lowest score being 880 and the highest being 2485. The split-half

reliability with Spearman-Brown correction was .719 for this task. As for distribution, both Z_{skew} and Z_{kurtosis} were less than the absolute value of 1.96; therefore, this variable was considered normally distributed, and no transformation was made.

Ospan. The Ospan measure drew on scores from 78 participants because the scores from two participants had to be excluded (see 5.7.1). The possible maximum for Opsan was 75. The mean from this sample was 45.24 (*SD* = 17.55), with the lowest score being zero and the highest 71. The reliability of this measure in terms of internal consistency according to Cronbach's alpha was .799. As for distribution, Z_{skew} was -2.03, with the absolute value greater than 1.96; therefore, this variable was considered negatively skewed, and transformation was applied (i.e., reflect and square root, and then reflect, 10-Sqrt (80-Ospan)).

NWR. The NWR score was the maximum number of phonemes each participant was able to recall from any given list. The possible maximum was 22. The mean from this sample was 13.36 (*SD* = 1.77), with the lowest being 9 and the highest 19. To establish the stability of this variable as a measure of phonological STM span, reliability was calculated by taking the Pearson correlation between the maximum and the top 2nd scores (i.e., the highest score across all 16 trials, and the second highest across all trials) for each participant. The correlation was .782. The distribution of this variable was normal, with both Z_{skew} and Z_{kurtosis} less than one. No transformation was used.

Pitch STM. The Pitch STM test generated two scores: the score for the control condition, and the score for the interference condition. These two scores were used as separate measures following Bowles et al. (2016), because it was shown in

their study that the two scores differentially predicted learning outcomes. The reliability in terms of internal consistency according to Cronbach's alpha was .858 for the control condition, i.e., the Pitch STM-control score, based on the original set of 48 items. Cronbach's alpha for the interference condition based on the original set of 48 items for that condition, however, was only .216, indicating very poor internal consistency. Item analysis was then run on the Pitch STM-interference score. The corrected item-total correlation was checked and items with negative item-total correlations or with small positive item-total correlations (i.e., less than .03) were excluded. This procedure resulted in an exclusion of 23 items, and the Cronbach's alpha of the remaining 25 items for the interference then reached .708. The percentage accuracy for the interference score was based on the retained 25 items.

As can be seen in Table 8, the mean for the Pitch STM-control scores was 83.65 (*SD* = 12.64), with the lowest being 52.08 and the highest 100. Comparatively, the mean for the Pitch STM-interference scores was lower, which was 68.55 (*SD* = 15.84), and with a larger range, i.e., the lowest being 20 and the highest 96. As for distribution, both variables were negatively skewed (the Z_{skew} for Pitch STM-control was -3.23, and the Z_{skew} for Pitch STM-interference was -2.14); therefore, transformations were used (i.e., reflect and square root, and then reflect).

CVMT. The CVMT task generated two scores: a *d*-prime score for the acquisition task (i.e., CVMT-acquisition), and a score in terms of the number correctly recalled for the delayed recognition task (i.e., CVMT-delayed). The split-half reliability with Spearman-Brown correction was .673 for CVMT-acquisition, and .660 for CVMT-delayed. The mean of the *d*-prime scores for CVMT-acquisition was

2.15 (*SD* = 0.54), with the lowest being 1.26, and the highest 3.71. The mean for CVMT-delayed was 4.98 (*SD* = 1.60), with the lowest being 1 and the highest 7. The distribution of CVMT-acquisition was positively skewed (Z_{skew} = 2.10), and CVMT-delayed was negatively skewed (Z_{skew} = -2.96). Therefore, transformations were applied as appropriate (i.e., square root for CVMT-acquisition; for CVMT-delayed, reflect, square root, and reflect).

SRT. The SRT score was the number of blocks (from block 3-8) participants showed learning as compared to the average learning effect across the latter six blocks based on the whole sample (see 5.7.5 for detail about the scoring procedure). The mean was 2.73 (*SD* = 1.58), with the lowest being 0 and the highest being 6). The split-half reliability with Spearman-Brown correction was .666, and the distribution seems to be normal, with both Z_{skew} and $Z_{kurtosis}$ within the cutoff range.

Table 8. *Descriptive Statistics of the Aptitude Measures*

	N	Possible Max	M	SD	Min	Max	Reliability
ShapeB	80	3,690	1625.75	415.40	880	2485	.719[a]
Ospan	78	75	45.24	17.55	0	71	.799[b]
NWR	80	22	13.36	1.77	9	19	.782[c]
PitchSTM_con	80	100	83.65	12.64	52.08	100	.858[b]
PitchSTM_int	80	100	68.55	15.84	20	96	.708[b]
CVMT_acq	80	--	2.15	0.54	1.26	3.71	.673[a]
CVMT_delayed	80	7	4.98	1.60	1	7	.660[a]
SRT	80	6	2.73	1.58	0	6	.666[a]
ProdTonalMem	80	1	0.28	0.27	0	.94	.922[b]

a. Split-half reliability with Spearman-Brown correction; b. Cronbach's alpha; c. correlation between the Max and Top 2nd scores

Productive tonal memory test. The productive tonal memory test score was the accuracy rate of the number of correctly sang tunes among a total of 18. The mean was 0.28 (*SD* = 0.27) with the lowest score being 0 and the highest .94. The reliability

in terms of internal consistency using Cronbach's alpha was .922. As for distribution, Z_{skew} was 1.80 and $Z_{kurtosis}$ was -2.03, and inverse transformation was applied (i.e., 1-1/(X+1)).

Table 9. *Distributions of the Aptitude Measures*

Aptitude Measures	Skewness			Kurtosis		
	Statistics	Std. Error	Z	Statistics	Std. Error	Z
ShapeB	0.038	0.269	0.14	-0.943	0.532	-1.77
Ospan	-0.553	0.272	-2.03	-0.395	0.538	-0.73
NWR	0.212	0.269	0.79	0.496	0.532	0.93
PitchSTM_con	-0.841	0.269	-3.13	-0.050	0.532	-0.09
PitchSTM_int	-0.575	0.269	-2.14	-0.077	0.532	-0.14
CVMT_acq	0.564	0.269	2.10	0.075	0.532	0.14
CVMT_delayed	-0.797	0.269	-2.96	-0.041	0.532	-0.08
SRT	0.390	0.269	1.45	-0.343	0.532	-0.64
ProdTonalMem	0.484	0.269	1.80	-1.079	0.532	-2.03

6.1.2 PCA on the aptitude measures

This section presents the results of a Principle Component Analysis (PCA) on the nine cognitive aptitude measures. PCA, an exploratory factor analytic technique, was used to validate whether the set of cognitive aptitude variables actually measured the theoretical constructs they were hypothesized to measure, i.e., Shapebuilder, Opsan, and NWR on WM, Pitch STM-control, Pitch STM-interference, and productive tonal memory on musical aptitude, CVMT-acquisition and CVMT-delayed on Declarative Memory (DM) ability, and SRT on Procedural Memory (PM) ability. In running the PCA and the correlational analysis, the aptitude scores with corrected distributions (i.e., transformed scores), if needed, were used.

Table 10 presents the intercorrelations between the nine cognitive measures. As can be seen, Ospan significantly correlated with Shapebuilder (r=.454, p<.001) and with NWR (r=.310, p=.006). Shapebuilder also significantly correlated with CVMT-acquisition (r=.291, p=.010), and CVMT-delayed (r=.336, p=.003). The musical ability test, i.e., the productive tonal memory measure, significantly correlated with Pitch STM-control (r=.544, p<.001) and Pitch STM-interference (r=.236, p=.037), with the correlation between Pitch STM-control and Pitch STM-interference being .247 (p=.029). In addition, CVMT-acquisition significantly correlated with CVMT-delayed (r=.449, p<.001). Interestingly, SRT correlated with CVMT-delayed (r=.274, p=.015).

Table 10. *Correlations between 9 Aptitude Measures (N=78)*

Measure		2	3	4	5	6	7	8	9
1.ShapeB	***r***	**.454****	**0.117**	**-0.116**	**0.051**	**.291****	**.336****	**0.070**	**-0.068**
	p	0.000	0.310	0.313	0.660	0.010	0.003	0.542	0.552
2.Ospan_	***r***		**.310****	**0.025**	**0.148**	**0.182**	**0.188**	**-0.086**	**-0.087**
refSqrt_ref	*p*		0.006	0.829	0.197	0.111	0.099	0.455	0.451
3.NWR	***r***			**0.065**	**0.008**	**-0.008**	**-0.042**	**0.150**	**-0.010**
	p			0.572	0.948	0.948	0.717	0.189	0.927
4.PitchSTM_	***r***				**.247***	**-0.009**	**-0.098**	**-0.004**	**.544****
Con_refSqrt_ref	*p*				0.029	0.936	0.394	0.976	0.000
5.PitchSTM_Int	***r***					**0.070**	**0.007**	**0.202**	**.236***
_refSqrt_ref	*p*					0.541	0.955	0.077	0.037
6.CVMT_Acq_	***r***						**.449****	**0.171**	**0.009**
sqrt	*p*						0.000	0.133	0.935
7.CVMT_delaye	***r***							**.274***	**-0.122**
d_refSqrt_ref	*p*							0.015	0.285
8.SRT	***r***								**0.148**
	p								0.195
9.ProdTonalMe	*r*								
m_inv_corrected	*p*								

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

A PCA was conducted (N_{subj}=78) on the nine aptitude measures. An oblique rotation method (Promax) was used, because the factors to be extracted were hypothesized to be correlated as all nine variables were conceptually related (they all measured aptitude). The analysis yielded four components with eigenvalues greater than 1.0 that explained 68.726% of the total variance. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .547 (greater than .500), and the Bartlett's test of sphericity was significant ($p < .001$) indicating that the correlation matrix differed significantly from zero. The first component had an eigenvalue of 2.089 and accounted for 23.215% of the variance; the second component had an eigenvalue of 1.772 and accounted for an additional 19.686% of the variance; the third component had an eigenvalue of 1.297 and accounted for an additional 14.413% of the variance; and the fourth component had an eigenvalue of 1.027 and accounted for an additional 11.412% of the variance.

Table 11 presents the component loadings of the nine variables on each of the four extracted components (the rotated Pattern Matrix). It showed that CVMT-acquisition and CVMT-delayed loaded strongly on the first component (.771 and .797, respectively), with Shapebuilder also loading quite substantially on this first component (.509). For the second component, it is clear that Pitch STM-control, Pitch STM-interference, and Productive Tonal Memory ability loaded strongly on this component (.837, .563, and .818), with the other variables having almost negligible loadings. The third component had strong loadings from Ospan (.836), NWR (.719) and Shapebuilder (.537), with negligible loadings from the other variables. With regard to the fourth component, only SRT loaded heavily on this component (.912),

with the second largest from NWR which was below .4. Except for Shapebuilder, which had almost equal loadings on Component 1 and Component 3, the other variables clearly loaded on a single component as expected. Shapebuilder, an established measure of WM, was expected to load on a component with Ospan; the reason why it also loaded quite heavily on Component 1 with the two CVMT measures was probably because both Shapebuilder and CVMT were visual tasks, and the CVMT obviously puts a burden on visual working memory. The distribution of the nine aptitude variables on the four components suggests that Component 1 represents DM ability, Component 2 musical aptitude, Component 3 WM, and Component 4 PM ability, which confirmed the proposed underlying structure of the cognitive aptitude measures in this study.

Table 11. *PCA Component Loadings (N=78)*

Variable	Component			
	1	2	3	4
ShapeB	**0.509**	-0.037	**0.537**	-0.117
Ospan_refSqrt_ref	0.180	0.077	**0.836**	-0.243
NWR	-0.351	-0.044	**0.719**	0.398
PitchSTM_Con_refSqrt_ref	-0.095	**0.837**	0.060	-0.102
PitchSTM_Int_refSqrt_ref	0.130	**0.563**	0.132	0.173
CVMT_Acq_sqrt	**0.771**	0.089	0.028	0.100
CVMT_delayed_refSqrt_ref	**0.797**	-0.111	0.000	0.244
SRT	0.274	0.070	-0.121	**0.912**
ProdTonalMem_inv_corrected	-0.025	**0.818**	-0.102	0.074

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.

6.1.3 Composite Aptitude Scores for Further Analyses

On the basis of the PCA results, three equally weighted composite scores were created: one for musical aptitude with the variables of Pitch STM-control, Pitch

STM-interference and Productive Tonal Memory combined, one for WM (in the narrow sense, i.e., complex WM) with Shapebuilder and Ospan combined, and one for DM ability with the two CVMT measures combined. A decision was made to not combine NWR with Shapebuilder and Ospan for complex WM, but to leave it as a separate measure of phonological STM, because phonological STM and complex WM are theoretically distinct constructs. The three composite scores were created by first converting each of the individual variables to *z*-scores and then taking the average of the *z*-scores of the variables for the composite. The scores for NWR and SRT were also converted to z-scores. The nine aptitude variables were then reduced to five, i.e., a composite score for musical aptitude, a composite score for WM, a composite score for DM ability, NWR as a measure of phonological STM, and SRT as a measure of PM ability. These five scores were then used as covariates in further hypothesis testing to answer the research questions.

The distributions of the three newly generated composite scores were checked: both Z_{skew} and $Z_{turtosis}$ were within the cutoff range (-1.96 to 1.96) for each composite score (see Table 12). As stated above, the original variables of NWR and SRT were normally distributed. Thus, all five aptitude covariates used in the further statistical testing followed a normal distribution. In addition, the four experimental groups did not differ on any of the five aptitude covariates (see Table 13 for a summary of the *F* test results), so the assumption of independence of the covariates and treatment was met. Table 14 presents the correlations between the five aptitudes' construct scores (covariates). Musical aptitude and PM ability (SRT) appeared to be very independent, and do not correlate with other aptitude scores. DM ability scores

(CVMT) correlated with WM scores (r=.324, p=.01); WM scores correlated with both CVMT and with NWR (r=.263, p=.03). The correlations were quite low and thus should not be of concern for multicollinearity in later multiple regression analyses.

Table 12. *Distribution of the Three Composite Aptitude Scores*

Composite Scores	Skewness			Kurtosis		
	Statistics	*SE*	*Z*	*Statistics*	*SE*	*Z*
WM_ZShapeBOspan	-0.470	0.272	-1.72	-0.470	0.538	-0.87
CVMT_ZAcqDelayed	-0.455	0.272	-1.67	0.017	0.538	0.03
Music_Z2PitchSTM1PTM	-0.139	0.272	-0.51	-0.966	0.538	-1.79

Table 13. *Group Differences on Aptitude Covariates to be used in hypothesis testing*

Aptitude Covariates	*df1*	*df2*	*F*	*p*
CVMT_ZAcqDelayed	3	64	0.224	0.879
WM_ZShapeBOspan	3	64	0.433	0.730
NWR_Z	3	64	1.140	0.340
Music_Z2PitchSTM1PTM	3	64	0.001	1.000
SRT_Z	3	64	0.605	0.614

Table 14. *Correlations between the 5 Aptitude Construct Scores (Covariates) (N=68)*

		WM_Z ShapeBOspan	NWR_Z	Music_Z 2PitchSTM1PTM	SRT_Z
CVMT_Z	***r***	**0.324****	**0.02**	**-0.08**	**0.17**
AcqDelayed	*p*	0.01	0.90	0.52	0.16
WM_Z	***r***		**.263***	**-0.04**	**-0.01**
ShapeBOspan	*p*		0.03	0.72	0.95
NWR_Z	***r***			**0.01**	**0.22**
	p			0.96	0.07
Music_Z	***r***				**0.17**
2PitchSTM1PTM	*p*				0.17
SRT_Z	*r*				
	p				

6.2 Language Learning Outcomes

This section presents the descriptive results of the outcome measures. The results of the accuracy outcome measures are presented first, followed by the RT outcome measures.

6.2.1 Accuracy measures

There were ten accuracy outcome measures, i.e., TS1_Pre_OWN_Tone_Acc for tone accuracy rate in the oral word naming task administered as pretest at the beginning of the first training session (TS1); TS1_Post_OPicN_4com_Acc for the accuracy rate of the four components of disyllabic words (including both segments and tone for each syllable) in the oral picture naming task administered at the end of TS1; TS2_Pre_OPicN_4com_Acc for TS2 pre-session quiz accuracy rate in oral picture naming; TS2_Post_OPicN_4com_Acc for TS2 post-session quiz accuracy rate in oral picture naming; TS3_Pre_OPicN_4com_Acc for TS3 pre-session quiz accuracy rate in oral picture naming; TS3_Post_OPicN_4com_Acc for TS3 post-session quiz accuracy rate in oral picture naming; D5_WPicN_4comp_Acc for the written picture naming accuracy rate on the retention test administered on Day 5; D5_OPicN_4com_Acc for the oral picture naming accuracy rate on the retention test administered on Day 5; D5_OWN_old_Tone_Acc for tone accuracy rate on old/practiced words in the oral word naming task administered on Day 5; and lastly D5_OWN_new_Tone_Acc for tone accuracy rate on new/generalization words in the oral word naming task administered on Day 5. Table 15 presents the reliability as measured by Cronbach's alpha for the list of the 10 accuracy outcome measures.

Table 15. *Reliability of the Accuracy Outcome Measures (N=70)*

No.	Outcome Accuracy Measures	Reliability (Cronbach's alpha)
1	TS1_Pre_OWN_Tone_Acc	.755
2	TS1_Post_OPicN_4com_Acc	.816
3	TS2_Pre_OPicN_4com_Acc	.887
4	TS2_Post_OPicN_4com_Acc	.814
5	TS3_Pre_OPicN_4com_Acc	.864
6	TS3_Post_OPicN_4com_Acc	.803
7	D5_WPicN_4comp_Acc	.864
8	D5_OPicN_4com_Acc	.808
9	D5_OWN_old_Tone_Acc	.843
10	D5_OWN_new_Tone_Acc	.953

Table 16 presents the means and standard deviations of each of the four groups (experimental conditions) on each of the ten accuracy outcome measures. The pretest in oral word naming of the to-be-learned target words was administered after the pre-training steps (i.e., after explicit instruction and the same amount of tone perception and production practice on monosyllabic words), and before disyllabic word training. The tone production accuracy rates were rather low in all four groups (*M*=.31 for Group A; *M*=.26 for Group B; *M*=.27 for Group C; *M*=.26 for Group D), which was expected because it was their first time reading disyllabic Mandarin words. A one-way ANOVA was conducted on the square root transformed scores of this variable (as the original distribution was positively skewed) with group as an independent variable. The equality of variances assumption was met, according to Levene's test (*p*=.761). The result showed that there were no significant differences between the four groups, F (3, 66) = .433, *p*=.730, indicating that the four groups started at the same level for tone production in disyllabic words.

Table 16. *Descriptive Statistics of the Accuracy Outcome Measures (in percentage correct) (N=70)*

Measures	Group A ISI-1d; RI-1w (n=18)		Group B ISI-1d; RI-4w (n=19)		Group C ISI-1w; RI-1w (n=16)		Group D ISI-1w; RI-4w (n=17)	
	M	SD	**M**	SD	**M**	SD	**M**	SD
Pretest in Oral Word Naming								
TS1_Pre_OWN_Tone_Acc	**0.31**	0.17	**0.26**	0.15	**0.27**	0.13	**0.26**	0.12
Oral Picture Naming across the Training Sessions								
TS1_Post_OPicN_4com_Acc	**0.45**	0.13	**0.47**	0.15	**0.46**	0.12	**0.44**	0.11
TS2_Pre_OPicN_4com_Acc	**0.37**	0.14	**0.38**	0.17	**0.25**	0.17	**0.25**	0.10
TS2_Post_OPicN_4com_Acc	**0.69**	0.11	**0.72**	0.15	**0.72**	0.11	**0.69**	0.08
TS3_Pre_OPicN_4com_Acc	**0.64**	0.13	**0.65**	0.17	**0.53**	0.16	**0.56**	0.10
TS3_Post_OPicN_4com_Acc	**0.82**	0.09	**0.83**	0.10	**0.81**	0.10	**0.83**	0.08
Retention Test after a delay								
D5_OPicN_4com_Acc	**0.71**	0.11	**0.55**	0.16	**0.69**	0.15	**0.62**	0.09
D5_WPicN_4comp_Acc	**0.80**	0.11	**0.61**	0.17	**0.82**	0.12	**0.71**	0.11
D5_OWN_old_Tone_Acc	**0.67**	0.23	**0.66**	0.20	**0.66**	0.20	**0.62**	0.12
D5_OWN_new_Tone_Acc	**0.64**	0.22	**0.59**	0.19	**0.62**	0.16	**0.52**	0.16

Note. For oral word naming tasks, the accuracy rates were based on the scores of the two *tone* components; for oral picture naming tasks, the accuracy rates were based on the scores of the four components (i.e., both the segments and the tone components for each of the two syllables of a word).

To keep track of learning within training sessions and forgetting between training session intervals, an oral picture naming task was administered at the end of TS1, and then the beginning and the end of both TS2 and TS3. See Table 16 for the

means and SDs of each of the four groups on oral picture naming performance at these five time points. Figure 5 graphically presents the development of oral picture naming performance of each of the groups across training sessions. It can be seen that by the end of TS1, the four groups reached almost the same level in oral picture naming, *M*=.45 for Group A; *M*=.47 for Group B; *M*=.46 for Group C; *M*=.44 for Group D. When they came back for the 2nd training session, either after a day or after a week, at the pre-session quiz all four groups performed at a lower level than they were at the end of TS1, due to memory decay. It is worth noting that the two massed practice groups (i.e., Groups A and B with an ISI of one day) seemed to have forgotten less than the two distributed groups (i.e., Groups C and D with an ISI of one week), *M*=.37 for Group A; *M*=.38 for Group B; *M*=.25 for Group C; *M*=.25 for Group D. After the 2nd training session, the four groups then seemed to have reached the same level by the end of TS2 in oral picture naming, *M*=.69 for Group A; *M*=.72 for Group B; *M*=.72 for Group C; *M*=.69 for Group D. After another ISI, at the pre-session quiz of TS3, again, degradation was observed, and the extent of degradation seems to be larger for the two distributed practice groups with longer ISI (1 week) than for the two massed practice groups with shorter ISI (1 day), *M*=.64 for Group A; *M*=.65 for Group B; *M*=.53 for Group C; *M*=.56 for Group D. After the 3rd training session, all four groups again reached almost the same level of performance in oral picture naming, *M*=.82 for Group A; *M*=.83 for Group B; *M*=.81 for Group C; *M*=.83 for Group D. The development of oral picture naming performance across the training sessions clearly shows the effectiveness of training, with the grand means across all

groups starting from .46 at the end of TS1, and then .71 by the end of TS2, and finally .82 by the end of TS3.

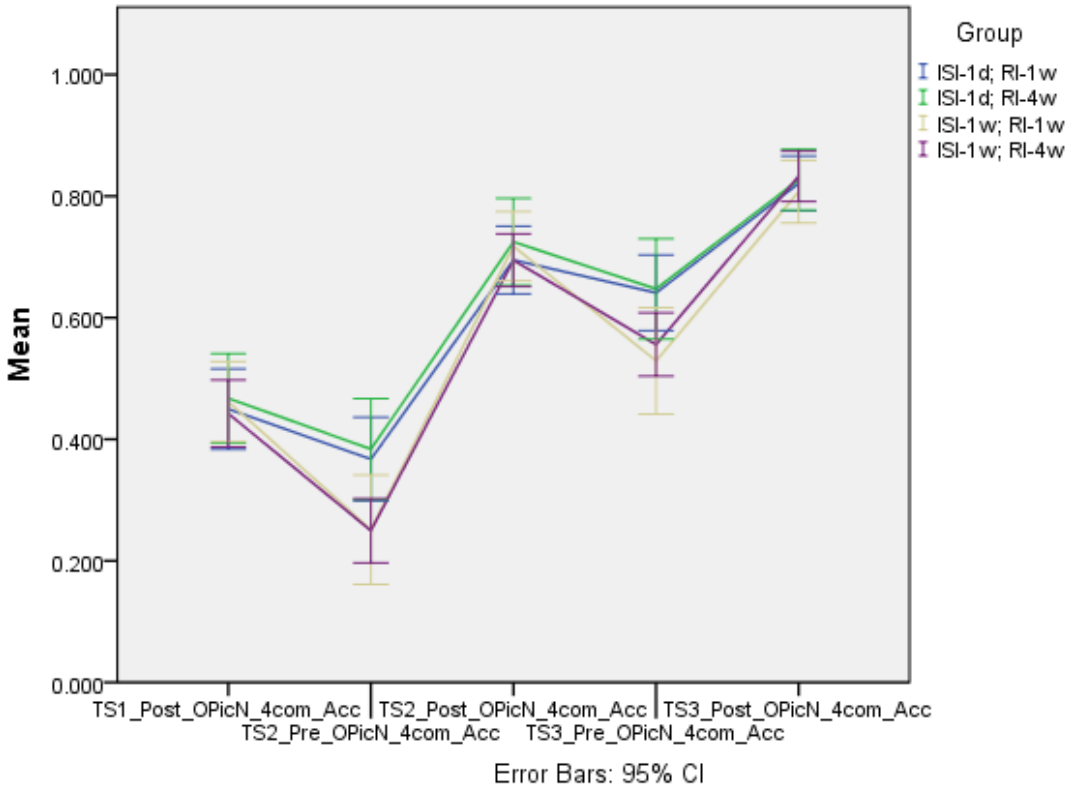


Figure 5. Development of Oral Picture Naming Performance across Training Sessions in the Four Experimental Groups

After an RI, i.e., after either one week or four weeks from the end of TS3, all participants came back for the last session in which the retention test was administered. See Table 16 for the means and SDs of the four groups on each of the four accuracy outcome measures from the retention test. Figure 6 presents the bar charts of each group on each measure on the same scale. For oral picture naming on Day 5, compared with oral picture naming performance at the end of the last training session (.82 in accuracy rate on average for all groups), memory delay was observed, and the extent of delay seemed to differ between groups. Descriptively, Group A

(massed practice with short-term 1-week RI) showed the least delay and performed at the highest level (*M*=.71), which was then followed by Group C (distributed practice with short-term 1-week RI) (*M*=.69). Group B (massed practice with long-term 4-week RI) performed at the lowest level (*M*=.55) showing the largest amount of delay, and Group D performed at a level in between (*M*=.62), higher than Group B and lower then Groups A and C. The accuracy rate drop in oral picture naming performance from the end of TS3 to the retention test on Day 5 was .11 for Group A, .12 for Group C, .21 for Group D, and .28 for Group B.

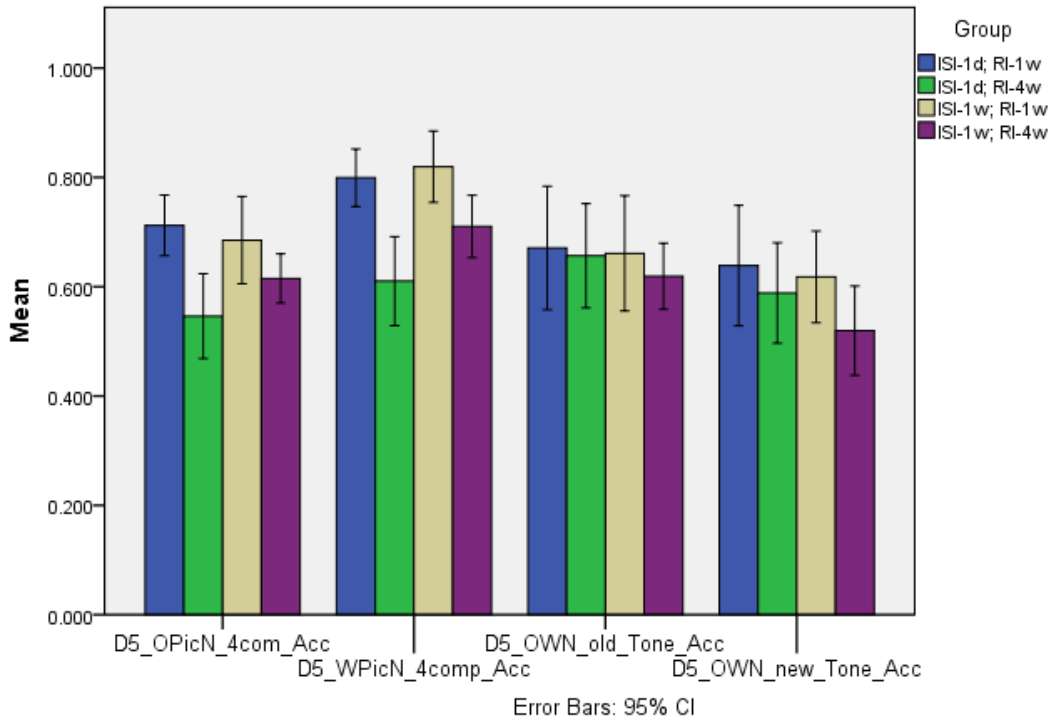


Figure 6. Means of the Four Accuracy Outcome Measures on the Retention Test Across Groups

As for written picture naming accuracy, a measure of the declarative component of Mandarin tonal word production (meaning-Pinyin associations), the

pattern of the results was similar to that of oral picture naming accuracy, except that all groups performed at a slightly higher level. Group A and Group C, i.e., the two groups with the short-term 1-week RI, performed at the highest level (*M*=.80 for Group A, and *M*=.82 for Group C); Group B, the group with short ISI (1 day) and long-term RI (4 weeks) performed at the lowest level (*M*=.61); Group D, the group with longer ISI (1 week) and long-term RI (4 weeks) performed at a level in between (*M*=.71), higher than Group B but lower than Groups A and C.

When it comes to retention performance in oral word naming, i.e., oral tonal production skill, the pattern seemed to be strikingly different (see Figure 6). Note that for oral *word* naming (reading aloud from Pinyin), the accuracy rates were based on the scores of the two *tone* components only because this task type focused on assessing *tone* production, whereas for oral *picture* naming tasks, the accuracy rates were based on the scores of the four components (i.e., both the segments and the tone components for each of the two syllables of a word) because participants needed to remember all four components of a disyllabic word in order to orally produce it. Although oral word naming is an easier task than oral picture naming, due to the difference in calculating the accuracy rates for the two task types, the accuracy rates for oral picture naming can be higher than the accuracy rates for oral word naming because tones were much harder to be pronounced correctly than segments for the participants in the study and higher segments accuracy rates can pull up the scores based on four components for the oral picture naming tasks.

All four groups seemed to perform at the same level in tone production accuracy on the old words, *M*=.67 for Group A, *M*=.66 for Group B, *M*=.66 for Group

C, and *M*=.62 for Group D. The 95% CIs almost paralleled among Groups A, B and C, and the CI for Group D fell within the CIs of the other three groups. In other words, irrespective of RI or ISI, oral tone production accuracy on practiced words were retained at the same level.

Although the same oral *word* naming task was not administered at the end of TS3 to make a direct comparison with oral word naming on Day 5 possible, in order to see whether the RIs resulted in any skill delay in oral tone production accuracy, the following additional analysis was conducted on the oral *picture* naming performance at the end of TS3. The component scores for segment accuracy and tone accuracy were calculated separately on oral picture naming at the end of TS3. See Table 16 for the group means. By the end of the third training session, all participants remembered the picture-Pinyin mappings of the target 20 words very well, which is reflected by the ceiling component scores for segments accuracy (the grand mean was .95 for four groups) in oral picture naming. An average error rate of .05 means that only 2 points were taken off from a total of 40 points allocated to the segments of 20 disyllabic words per participant. Among the 2 points that were taken, most of the times it was because their pronunciation on segments was not native-like and only in a few cases, it was because participants did not remember the word for the picture. Therefore, the component score for *tone* accuracy in this oral *picture* naming task can be considered a good approximation of *tone* accuracy in oral *word* naming. See Table 17 for the mean tone accuracy rates in oral picture naming at the end of TS3 for each group. The mean accuracy rate were .69 for Group A, .71 for Group B, .66 for Group C, and .71 for Group D. Recall that the means of tone accuracy in oral word naming on Day 5

were .67 for Group A, .66 for Group B, .66 for Group C and .62 for Group D. Thus, the skill decay reflected in tone accuracy rate drop after an RI was only .02 and .00 for Group A and Group C (the two groups with a 1-week RI) and .05 and .09 for Group B and Group D (the two groups with a 4-week RI). That is, skill delay in oral tone production accuracy was almost negligible with a 1-week delay, and very small with a 4-week delay.

Table 17. *The Segments and Tone Component Accuracy Rate in TS3_Post_OPicN*

TS3_Post_OPicN	Seg Acc		Tone Acc	
	M	SD	**M**	SD
ISI-1d; RI-1w (n=18)	**0.95**	0.05	**0.69**	0.16
ISI-1d; RI-4w (n=19)	**0.94**	0.05	**0.71**	0.18
ISI-1w; RI-1w (n=16)	**0.95**	0.04	**0.66**	0.17
ISI-1w; RI-4w (n=17)	**0.96**	0.03	**0.71**	0.15

With regard to oral word naming performance on new words, the scores were generally lower compared to the oral word naming scores on old/practiced words, *M*=.64 for Group A, .59 for Group B, .62 for Group C, and .52 for Group D. The group means on new words were not as even as the group means on the old words, but the differences seemed to be small and the CIs all overlap across the four groups. Descriptively, the mean differences between oral word naming on *new* words and oral word naming on *old* words seemed to be larger for Groups B and D (.07 and .10 respectively) than for Groups A and C (.03, .04 respectively).

6.2.2 RT measures

As RT was intended to be used as a measure of automaticity, only RTs for the responses with correct tone[4] production (i.e., getting 2 out of 2 for tones in a disyllabic word) were included as valid data points for further RT analysis. Although RTs were recorded in all oral production tasks, including oral word naming and oral picture naming administered during the training sessions and on the retention test, only the RTs at the later stages, i.e., on the immediate post-test at the end of TS3 and on the delayed retention test on Day 5 were calculated and used as RT outcome measures, because too few valid RT data points could be included (due to high error rates in tone production) at the earlier stages to generate reliable RT measures. In addition to responses with incorrect tones, the RTs for falsely triggered items were also excluded (see 5.6.1 for detail about false trigger). See Table 18 for the amount of RTs excluded due to each of the two excluding procedures on the four RT outcome measures. The total amount of RT data excluded for TS3_Post_OPicN_RT was 45.1%. For D5_OPicN_RT, 65.6% were excluded. For D5_OWN_old_RT, 49.9% were excluded, and for D5_OWN_new_RT, 59.2% were excluded.

Table 18. *RT Data Cleaning Procedure (N=70)*

	Total	Excluded due to incorrect tones	Additional due to False Triggers	Total excluded	Percentage excluded
TS3_Post_OPicN_RT	1400	625	6	631	0.451
D5_OPicN_RT	1400	909	10	919	0.656
D5_OWN_old_RT	1400	697	2	699	0.499
D5_OWN_new_RT	4480	2620	31	2651	0.592

[4] The criterion was NOT set to require getting 4 out of 4 for each disyllabic word (1 point for each of the four components, i.e., segments and tone for each syllable), but was set to require getting 2 out of 2 for the *tone* components because this study focused on tone production. A decision was made to ignore inaccuracy in segment pronunciation but to rely only on tone production correctness for inclusion of RTs.

After the above two exclusion procedures, the means of the valid RTs for each participant on each of the four RT outcome measures were computed, together with the number of items included for calculating the mean, the SD for the mean, and the coefficient of variation (CV), i.e., SD/M, as an index of stability/restructuring in automatization (Segalowitz & Segalowitz, 1993). It was observed that the RT means for a few participants were based on a very limited number of items. A decision was then made to exclude those participant RT means that were based on less than three items. This procedure resulted in an exclusion of 2 participant means (1 in Group B and 1 in Group C) for the TS3_Post_OPicN_RT measure, an exclusion of 7 participant means (2 in Group B, 4 in Group C and 1 in Group D) for the D5_OPicN_RT measure, an exclusion of 4 participant means (1 in Group A, 2 in Group B, and 1 in Group C) for the D5_OWN_old_RT measure, and an exclusion of 1 participant (in Group B) for the D5_OWN_new_RT measure.

For the EngOWN_RT measure, 8 RTs were excluded due to false triggers, and additional 32 outliers (i.e., 3 SDs above or below each participant's mean RT) were excluded. These procedures resulted in an exclusion of 1.9% of RT data for this measure.

Split-half reliability, based on the RT data points after the cleaning procedures, was calculated for the four RT outcome measures. The Split-half *R* was calculated based on the formula $R=2r/(r+1)$, in which *r* is the Pearson correlation between participants' RT means on odd items and those on the even items in each test. See Table 19 for the reliability of the RT measures.

Table 19. *Reliability of the RT Measures (N=70)*

RT measures	Reliability (Split-half *R*)
Eng_OWN_RT	.990
TS3_Post_OPicN_RT	.762
D5_OPicN_RT	.631
D5_OWN_old_RT	.885
D5_OWN_new_RT	.952

Tables 20-23 present the average RT means, SDs and CVs of the four experimental groups on each of the four RT outcome measures. The number of participants and average number of items included for calculating these indexes are also presented in the tables. Furthermore, the correlation between the mean RT and CV was calculated for each group on each of the outcome measures and presented in the tables. According to Segalowitz and Segalowitz (1993), a positive correlation between RT and CV indicates automatization in the narrow sense of *restructuring*, i.e., "a qualitative change due to the dropping out of some processing components or the modularization of processing" (p. 374).

Starting from the RTs for oral picture naming at the end of TS3 (See Table 20), the RT means were 1848 ms (*SD* = 792 ms) for Group A, 1744 ms (*SD* = 744 ms) for Group B, 1681 ms (*SD* = 712 ms) for Group C, and 1797 ms (*SD* = 727 ms) for Group D, with comparable SDs or variation across groups. The CVs were quite small in all groups, ranging from .38 to .41. The RT-CV correlations were .459 (p=.055) in Group A, .782 (p<.001) in Group B, .555 (p=.032) in Group C, and .600 (p=.011) in Group D, all positive and statistically significant (with the only exception for Group A with the correlation marginally significant), suggesting that

automatization in terms of restructuring/stability had occurred by the end of TS3 in all groups.

Table 20. *Means, SDs, CVs, and RT-CV correlations of the four groups on TS3_Post_OPicN_RT (in milliseconds)*

TS3_Post_ OPicN_RT	n_{subj}	n_{item}	RT_Mean	RT_SD	CV	RT-CV correlation	*p*
Group A	18	11.17	1848.06	791.52	0.41	.459	.055
Group B	18	12.06	1744.17	744.32	0.38	.782	<.001
Group C	15	10.93	1680.51	711.66	0.39	.555	.032
Group D	17	10.82	1797.03	727.19	0.38	.600	.011

As for oral picture naming on Day 5 (see Table 21), after a retention interval, participants' RTs on the same task became much slower and demonstrated much larger variation across all groups (see also Figure 7): the RT means were 3222 ms (*SD* = 1491 ms) for Group A, 4611 ms (*SD* = 2781 ms) for Group B, and 2603 ms (*SD* = 1560 ms) for Group C, and 2854 ms (*SD* = 1303 ms) for Group D. The CVs also became larger on Day 5 than at the end of TS3, ranging from .42 to .59. Among the four groups, the mean RTs were faster in the two distributed practice groups (Groups C and D) than the two massed practice groups (Groups A and B); in addition, the groups with longer RI (Groups B and D, ISI being the same) responded slower than the groups with shorter RI (Group A and Group C, ISI being the same). Group B, massed practice group with long-term 4-week RI, responded the slowest with the largest variation, and Group C, distributed practice group with short-term 1-week RI, responded the fastest. The RT-CV correlation was not significant in Group A (*r*=.203, *p*=.418) or Group B (*r*=.214, *p*=.410), suggesting that the level of automaticity was lost in oral picture naming after a delay of 1 week or 4 weeks for the two massed

practice groups with a daily training schedule. The RT-CV correlation became marginal in Group D (r=.480, p=.060), suggesting that the group with a weekly training schedule had a reduction of automaticity in oral picture naming after a 4-week delay. The RT-CV correlation only remained significant in Group C (r=.837, p=.001), suggesting that the group with a weekly training schedule retained a certain level of automaticity after a 1-week delay.

Table 21. *Means, SDs, CVs, and RT-CV correlations of the four groups on D5_OPicN_RT (in milliseconds)*

D5_OPicN_RT	n_{subj}	n_{item}	RT_Mean	RT_SD	CV	RT-CV correlation	*p*
Group A	18	8.17	3222.07	1491.43	0.45	.203	.418
Group B	17	6.88	4611.35	2780.61	0.59	.214	.410
Group C	12	9.42	2603.36	1560.38	0.51	.837	.001
Group D	16	6.06	2854.41	1302.93	0.42	.480	.060

When it comes to oral word naming on old/practiced words on Day 5, the task that removed the declarative component of picture-Pinyin mappings compared with oral picture naming, the RTs were much faster than the RTs in oral picture naming on Day 5. See Table 22 and Figure 8. The RT means were comparable among the four groups with relatively small and comparable variations. The CVs were the smallest on this RT outcome measure, ranging from .27 to .34, indicating a relatively high level of automaticity on this task. The RT-CV correlations were all positive and statistically significant across the four groups, suggesting a high level of automaticity and stability in oral tone production in all four groups.

Table 22. *Means, SDs, CVs, and RT-CV correlations of the four groups on D5_OWN_old_RT (in milliseconds)*

D5_OWN_old_RT	n_{subj}	n_{item}	RT_Mean	RT_SD	CV	RT-CV correlation	*p*
Group A	17	11.29	1470.39	481.29	0.30	.663	.004
Group B	17	11.00	1538.79	559.83	0.34	.545	.024
Group C	15	11.20	1432.92	467.96	0.30	.801	<.001
Group D	17	8.71	1317.60	398.22	0.27	.839	<.001

Table 23. *Means, SDs, CVs, and RT-CV correlations of the four groups on D5_OWN_new_RT (in milliseconds)*

D5_OWN_new_RT	n_{subj}	n_{item}	RT_Mean	RT_SD	CV	RT-CV correlation	*p*
Group A	18	29.83	2421.97	845.75	0.33	.429	.075
Group B	18	26.89	2593.82	911.63	0.33	.545	.019
Group C	16	27.63	2503.89	847.17	0.30	.691	.003
Group D	17	21.24	2329.39	780.39	0.32	.381	.131

With regards to oral word naming on new words (see Table 23 and Figure 7), compared to oral word naming on old words, the RTs for all four groups were slower, with larger variations (see SDs). The CVs seemed to be within the similar range, i.e., from .30 to .33, with those for oral word naming on old words (.27 to .34). The RT-CV correlation was positive and remained significant in Group B (r=.545, p=.019) and Group C (r=.691, p=.003); however, the correlation became nonsignificant in Group D (r=.381, p=.131) and Group A (r=.429, p=.075).

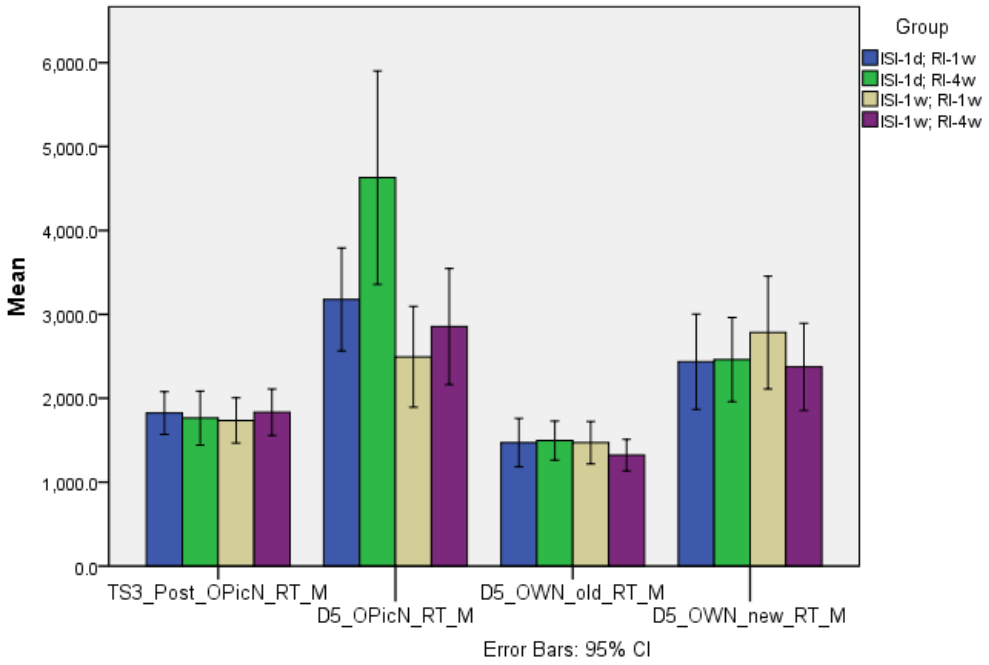


Figure 7. Means of the Four RT Outcome Measures Across Groups

Taken together, it seemed that automatization in terms of restructuring or stability had occurred in oral picture naming by the end of TS3, which was demonstrated by the low RT means, small variations, small CVs, and positive RT-CV correlations across four groups. In addition, a certain level of automaticity seemed to be retained in oral *word* naming (oral tone production) in old/practiced words on Day 5, which was also demonstrated by the low RT means, small variations, small CVs, and positive RT-CV correlations across four groups. For oral picture naming on Day 5, a task that requires the retention of a large portion of declarative knowledge (i.e., picture-Pinyin mappings), the level of automaticity retained seems to depend on ISI and RI. For oral word naming on new words, a task that requires synthesizing new combinations of segments with tones, the level of production automaticity also seems to depend on the groups, i.e., ISI or RI.

6.2.3 Distributions of all outcome measures for further hypothesis testing

The distributions of all language learning outcome measures (including accuracy rate and RT) are presented in Table 24. The distributions of all accuracy learning outcome measures look normal (with *Z* values less than 2 or marginal) except TS1_Pre_OWN_Tone_Acc, which is not literally a learning *outcome* measure, but a pre-test (not surprising that its distribution was positively skewed because they had not started to learn). Square root transformation was applied to this accuracy variable, only. For all other nine accuracy variables, the original scores were used for hypothesis testing. As for the RT variables, four of the five were positively skewed; therefore, log transformation was applied to all five RT variables. The log transformed RT variables were used for hypothesis testing.

Table 24. *Distributions of All Learning Outcome Measures*

Outcome Measures	N	Sknewness			Kurtosis		
		Statistics	*SE*	*Z*	*Statistics*	*SE*	*Z*
TS1_Pre_OWN_Tone_Acc	68	1.145	0.291	3.93	1.988	0.574	3.46
TS1_Post_OPicN_4com_Acc	68	0.058	0.291	0.20	-0.647	0.574	-1.13
TS2_Pre_OPicN_4com_Acc	68	0.249	0.291	0.86	-0.836	0.574	-1.46
TS2_Post_OPicN_4com_Acc	68	0.393	0.291	1.35	-0.676	0.574	-1.18
TS3_Pre_OPicN_4com_Acc	68	0.143	0.291	0.49	-0.402	0.574	-0.70
TS3_Post_OPicN_4com_Acc	68	-0.411	0.291	-1.41	-0.304	0.574	-0.53
D5_WPicN_4comp_Acc	68	-0.494	0.291	-1.70	-0.165	0.574	-0.29
D5_OPicN_4com_Acc	68	-0.122	0.291	-0.42	0.120	0.574	0.21
D5_OWN_old_Tone_Acc	68	-0.585	0.291	-2.01	0.089	0.574	0.16
D5_OWN_new_Tone_Acc	68	-0.193	0.291	-0.66	-0.433	0.574	-0.75
EngOWN_RT_M*	68	0.758	0.291	2.60	0.930	0.574	1.62
TS3_Post_OPicN_RT_M	66	0.213	0.295	0.72	-0.416	0.582	-0.71
D5_OPicN_RT_M	61	1.672	0.306	5.46	3.642	0.604	6.03
D5_OWN_old_RT_M	64	1.371	0.299	4.59	3.459	0.590	5.86
D5_OWN_new_RT_M	67	0.781	0.293	2.67	0.014	0.578	0.02

*EnglishOWN_RT_M is the only measure listed here that was not an outcome measure.

6.3 Effects of ISI on L2 learning across training sessions

The descriptive results of the four experimental groups on the pre- and post-session quizzes in oral picture naming administered at the beginning and end of the training sessions were reported in section 6.2. When we look at outcome performance across the training sessions, among the two independent variables (ISI and RI), only ISI can possibly have an effect on performance at the pre- or post- session quizzes across the training sessions because RI is the interval between the end of the last training session and the retention test, and thus can only exert effects on retention. In addition, ISI can only play a role on performances after an ISI has happened; therefore, performance on TS1 can not be affected by ISI, and ISI can only start to have an effect from the pre-session quiz on TS2 and thereafter.

The first research question regards the effects of ISI on L2 learning across training sessions. As this study has an experimental design with participants randomly assigned to the four experimental conditions, as no participants had any prior knowledge about Mandarin tones or any of the 20 target words, and as all participants went through the same training sessions including the first, by the end of TS1 all experimental groups were expected and also turned out to have achieved the same level of performance in the post-session quiz in oral picture naming, F (3, 66) = 0.134, p =.940 (see Table 15 for the descriptives of the four groups). A reminder that all four groups also performed at the same level in the pretest at the beginning of TS1 (p=.730) (see 6.2, Table 15). As all groups started from zero and achieved the same level of performance by the end of TS1 before ISI started to exert an effect, a comparison of different ISI groups on performances in the pre- and post-session

quizzes of TS2 and TS3 can demonstrate the effects of ISI on L2 learning (higher outcome performances, more learning/improvement).

To assess the effects of ISI, as RI has not started to play a role during the training phase, and all four experimental groups differed only on ISI, the four groups were then collapsed into two bigger groups, i.e., the 1-day ISI group (Groups A and B) and the 7-day ISI group (Groups C and D) for further statistical testing. Again, the two ISI groups (1-day vs 7-days) did not differ on the pretest in oral word naming accuracy at the beginning of TS1, F (1, 68) = .051, p =.822, nor did they differ on the post-session quiz in oral picture naming accuracy at the end of TS1, F (1, 68) = 0.052, p = .821. Figure 8 presents the means and 95% CIs of the ISI-1day group and the ISI-1week group on oral picture naming accuracy at the five time points across the training sessions.

To test whether the 1-day ISI group outperforms the 7-day ISI group across the 2nd and 3rd training sessions (at both the pre- and post- session quizzes) in learning oral Mandarin tonal word production (Hypothesis 1a), a two-way Repeated Measures ANOVA was conducted, with ISI as the between-subjects variable, and Time (TS2 pre-session quiz, TS2 post-session quiz, TS3 pre-session quiz, and TS3 post-session quiz) as a within-subjects variable on the oral picture naming accuracy performances at the four time points. According to Mauchly's test, the assumption of sphericity was violated for the main effects of Time, χ^2 (5) = 32.816, p < .001; therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε=.761). Levene's test shows that the equality of variance assumption was met for TS2_Pre (p=.406), TS3_Pre (p=.695), and TS3_Post (p=.630), indicating equal variability

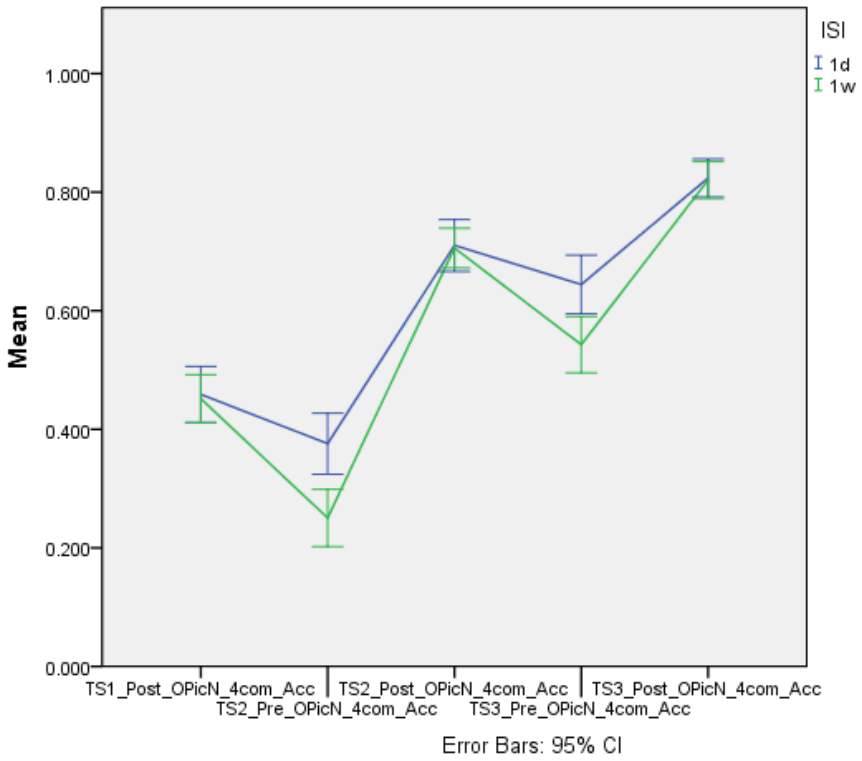


Figure 8. Development of Oral Picture Naming Performance across Training Sessions in the ISI-1day group and ISI-1week group

across groups on these variables, but was not met for TS2_Post (p = .011) indicating unequal variability across groups for this variable. However, for TS2_Post, the larger SD (.13) was less than two times the smaller SD (.09). Therefore, ANOVA was considered robust. The results show that there was a significant main effect of Time, F (2.282, 155.167) = 592.694, p < .001, η_p^2 = .897, and ISI, F (1, 68) = 5.156, p=.026, η_p^2=.070. More importantly, the interaction between Time and ISI was also significant, F (2.282, 155.167) = 12.713, p<.001, η_p^2=.158. The simple main effects of ISI at each level of Time were then tested using Bonferroni correction for multiple comparisons. The pairwise comparisons showed that the 1-day ISI group significantly outperformed the 1-week ISI group on the pre-session quiz at the beginning of TS2

(TS2_Pre_OPicN_4com_Acc), $F(1, 68)=12.842$, $p=.001$, $\eta_p^2=.159$, and on the pre-session quiz at the beginning of TS3 (TS3_Pre_OPicN_4com_Acc), $F(1, 68)=8.896$, $p=.004$, $\eta_p^2=.116$; however, the two ISI groups did not differ on the post-session quiz at the end of TS2 (TS2_Post_OPicN_4com_Acc), $F(1,68)=0.026$, $p=.873$, $\eta_p^2<.001$, or on the post-session quiz at the end of TS3 (TS3_Post_OPicN_4com_Acc), $F(1, 68)=0.025$, $p=.874$, $\eta_p^2<.001$. Table 25 summarizes these results reported.

Table 25. *Summary of results from Repeated Measures ANOVA for RQ1*

	RQ1. RM ANOVA		
	F	*p*	η_p^2
Time	592.694	0.000	0.897
Time*ISI	12.713	0.000	0.158
ISI	5.156	0.026	0.070
The simple effects of ISI on each level of Time			
Time1: TS2_Pre	12.842	0.001	0.159
Time2: TS2_Post	0.026	0.873	0.000
Time3: TS3_Pre	8.896	0.004	0.116
Time4: TS3_Post	0.025	0.874	0.000

Even though the two ISI groups did not differ on their exit performance in oral picture naming at the end of TS1, before ISI started to exert an effect, a repeated-measures ANCOVA was conducted using TS1_Post_OPicN_4com_Acc as a covariate to control any preexisting differences, and again with ISI as a between-subjects variable and Time as a within-subjects variable. Again, Mauchly's test indicated that the assumption of sphericity was violated, $\chi^2(5) = 26.982$, $p < .001$; therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon=.787$). The assumption of equality of error variances (Levene's test) was met for TS2_Post ($p=.096$), TS3_Pre ($p=.531$), and TS3_Post ($p=.866$), but not for

TS2_Pre (p=.018). For TS2_Pre, the larger SD (.15) was less than two times the smaller SD (.14). Therefore, the repeated measures ANCOVA was considered robust.

Table 26 summarizes the results from this analysis. Again, there was a significant main effect of Time, ISI, and the covariate, i.e., TS1_Post_OPicN_4com_Acc. The interactions between Time and the covariate and between Time and ISI were significant. As only the interaction between Time and ISI was of interest to the present study, further comparisons were only made on the effects of ISI at each level of Time. The same pattern of results were found with larger effect sizes on the effects of ISI on the pre-session performances at the beginning of TS2 and TS3, i.e., the 1-day ISI group performed much better than the 1-week ISI group on the pre-session quizzes of TS2 and TS3. However, the two groups' performances at the end of TS2 and TS3 were at the same level.

Table 26. *Summary of results from Repeated Measures ANCOVA for RQ1*

	RQ1. RM ANCOVA		
	F	p	η_p^2
Time	104.480	0.000	0.609
Time*ISI	14.150	0.000	0.174
ISI	12.858	0.001	0.161
TS1_Post_OPicN_4com_Acc	134.348	0.000	0.667
Time*TS1_Post_OPicN_4com_Acc	12.163	0.000	0.154
The simple effects of ISI on each level of Time			
Time1: TS2_Pre	30.669	0.000	0.314
Time2: TS2_Post	0.000	0.995	0.000
Time3: TS3_Pre	16.897	0.000	0.201
Time4: TS3_Post	0.002	0.964	0.000

Hypothesis 1a stated that the 1-day ISI group were expected to outperform the 1-week ISI group on oral picture naming accuracy across the board from TS2 pre-

session quiz, TS2 post-session quiz, to TS3 pre-session quiz and TS3 post-session quiz. However, this hypothesis was NOT confirmed. Instead, ISI only had an effect on the *pre*-session quizzes at the beginning of the subsequent two training sessions, but did not show effects on the post-session performances of the following training sessions. These results will be discussed in the discussion section.

To test Hypothesis 1b regarding whether the daily practice group would respond faster than the weekly practice group in oral picture naming at the end of the last training session, a one-way ANOVA was conducted with ISI as the independent variable and the log-transformed TS3_Post_OPicN_RT variable as the dependent variable. Levene's test suggests that the assumption of homogeneity of variance across groups was met (p=.599). The results showed that the main effect of ISI was not statistically significant, F (1, 66) = 0.065, p=.799, η_p^2=.001, suggesting that there was little difference between the two ISI groups on oral picture naming RT for items that they correctly produced tones. Hypothesis 1b was thus disconfirmed.

6.4 Effects of Aptitudes at different stages of L2 learning across training sessions

The RQ2 asks what roles cognitive aptitudes play at different stages of learning across the training sessions when the effect of ISI is controlled (if participants have gone through an ISI by the time of testing). At the early stage, i.e., at the end of first training session (TS1), WM, DM ability, PSTM and musical aptitude were hypothesized to play a facilitative role in oral picture naming accuracy performance. At the later stage, i.e., the end of the last training session (TS3), musical aptitude was still expected to play a facilitative role; in addition, PM ability was expected to play a facilitative role in oral picture naming accuracy and oral picture

naming RT when L1 word naming RT is controlled. In the following, results testing these two sets of hypotheses are presented in order.

6.4.1 Aptitude at Early Stage of Learning

To test whether WM capacity, DM ability, PSTM and musical aptitude play a facilitative role at the beginning stage of learning (Hypotheses 2a-i, ii, iii, iv), four simple linear regressions (Models 1-4) were conducted, by regressing the oral picture naming accuracy scores at the end of the first training session on each of the four aptitude construct scores, respectively. These single-predictor analyses were conducted to show how each predictor is independently related to the outcome measure in isolation, so as to avoid interpretive difficulties introduced by correlations between the predictors (see Table 14 for the correlation matrix between the five aptitude construct scores).

Results from the four simple linear regressions show that (1) DM ability (CVMT_ZAcqDelayed) was not a significant predictor on oral picture naming accuracy at the end of TS1, R^2=.034, F (1, 66)=2.350, p=.130; (2) WM capacity (WM_ZShapeBOspan) was not a significant predictor on TS1_Post_OPicN_4com_Acc, R^2=.021, F (1, 66)=1.385, p=.243; (3) PSTM (NWR_Z) was not a significant predictor either, R^2<.001, F (1, 66)=0.019, p=.892, and finally (4) Musical aptitude (Music_Z2PitchSTM1PTM) was not a significant predictor, R^2<.001, F (1, 66)=0.019, p=.891. In short, none of the four hypothesized aptitudes turned out to be an independent predictor of learning outcome performance in oral picture naming accuracy at the end of TS1. Table 27 presents a summary of the parameter estimates of the aptitudes in these single-predictor models.

Table 27. *Parameter Estimates of Aptitudes in Single-Predictor Analyses on TS1_Post_OPicN_4com_Acc*

		TS1_Post_OPicN_4com_Acc				
	Predictor	*B*	*SEB*	*t*	*p*	η_p^2
Model 1	CVMT_ZAcqDelayed	0.027	0.018	1.533	0.130	0.034
Model 2	WM_ZShapeBOspan	0.021	0.018	1.177	0.243	0.021
Model 3	NWR_Z	-0.002	0.016	-0.137	0.892	0.000
Model 4	Music_Z2PitchSTM1PTM	0.003	0.020	0.137	0.891	0.000

As the finding that none of the hypothesized aptitudes turned out to be significant predictor of outcome performance at the end of TS1 was rather puzzling, a follow-up analysis was conducted to see whether pretest performance predicts learning outcome at the end of TS1. Model 5 was conducted by regressing TS1_Post_OPicN_4com_Acc on TS1_Pre_OWN_Tone_Acc_sqrt, and found that pretest performance in tone production accuracy was a significant predictor of learning outcome at the end of TS1, R^2=.090, F (1, 66)=6.533, p=.013. See Table 28 for the parameter estimates of TS1_Pre_OWN_Tone_Acc_sqrt on TS1_Post_OPicN_4com_Acc; the positive coefficient (B=.272) shows that participants who performed better in the pretest in oral word naming accuracy achieved better learning outcome in oral picture naming accuracy by the end of TS1.

As there has been research (Li & DeKeyser, in press) showing that musical aptitude predicts Mandarin tone production accuracy in single syllable words, an additional analysis was run to see whether musical aptitude predicts tone production accuracy in disyllabic words in the pretest in this study. Model 6 was run by regressing TS1_Pre_OWN_Tone_Acc_sqrt on Music_Z2PitchSTM1PTM, and found that musical aptitude is a significant predictor of tone production accuracy in

disyllabic words in the pretest, R^2=.085, F (1, 66) = 6.125, p = .016 (See Table 29 for the parameter estimates). The positive coefficient (B=.053) suggests that participants with higher musical aptitude performed better in oral tone production in disyllabic words in the pretest.

Table 28. *Parameter Estimates of Pretest performance in Oral Word Naming on TS1_Post_OPicN_4com_Acc*

		TS1_Post_OPicN_4com_Acc				
	Predictor	*B*	*SEB*	*t*	*p*	η_p^2
Model 5	TS1_Pre_OWN _Tone_Acc_sqrt	0.272	.107	2.556	.013	.090

Table 29. *Parameter Estimates of Musical Aptitude on Pretest Performance in Oral Word Naming*

		TS1_Pre_OWN_Tone_Acc_sqrt				
	Predictor	*B*	*SEB*	*t*	*p*	η_p^2
Model 6	Music_Z2PitchSTM1PTM	0.053	.022	2.475	.016	.085

Thus, it seems that the relationship between musical aptitude and the learning outcome in oral picture naming accuracy at the end of TS1 was mediated by pre-training performance in oral tone production accuracy in oral word naming. To formally test this mediation hypothesis, mediation analyses were conducted using the bootstrapping method with bias-corrected confidence estimates (Preacher & Hayes, 2008). The 95% confidence interval of the indirect effects was obtained with 5000 bootstrap resamples (Preacher & Hayes, 2008). Results of the mediation analysis confirmed the mediating role of pre-training performance in tone production accuracy in the relation between musical aptitude and the outcome performance in oral picture

naming accuracy at the end of TS1 (*B* = .016; *CI* = .001 to .042). The fact that musical aptitude, the predictor, was significantly related to pretest performance in tone production accuracy, and pretest performance, the mediator, was significantly associated with the outcome performance, together with the finding of no direct effect of the predictor on the outcome, suggests a complete mediation effect.

To summarize, in response to RQ2-a, the learning outcome at early stage (oral picture naming accuracy at the end of TS1) was not directly predicted by WM capacity, DM ability, PSTM, or musical aptitude; instead, it was predicted by pretest performance in tone production accuracy in disyllabic words, which was then further predicted by musical aptitude. Pretest performance in tone production accuracy mediated the relationship between musical aptitude and the learning outcome at the end of the first training session in oral picture naming accuracy.

Appendix L presents the correlations between the five aptitude construct scores and the pre- and post-session quiz performance at the first training session.

6.4.2 Aptitude at Later Stages of Learning

RQ2b concerns the roles of cognitive aptitude at a later stage of learning. A later stage was operationalized as the outcome performance at the end of the last training session (TS3). Appendix M presents correlations between the five aptitude construct scores and the post-session quiz performance (in terms of both accuracy and RT) on TS3 across the two ISI groups (1-day ISI vs. 1-w ISI), respectively.

To test whether musical aptitude or PM ability plays a facilitative role in oral picture naming accuracy by the end of the last training session (TS3) (Hypotheses 2b-i, ii), separate ANCOVAs were conducted with ISI as an independent variable, and

musical aptitude or PM ability as covariates, on post-session quiz performance in oral picture naming accuracy at the end of TS3. ISI was kept in all ANCOVA models here in order to control any effect of ISI on outcome performance. ANCOVAs were first run with one aptitude score at a time as single covariates (without interactions); additional ANCOVAs were then run to allow the covariate to interact with the experimental factor to see whether there was an Aptitude-Treatment Interaction (ATI), i.e., in this case, whether any of the aptitudes interacted with ISI. Levene's test was conducted in each ANCOVA modeling, and the *p* values were all larger than .500, suggesting equality of error variances across groups. In addition, the two ISI groups (1-day ISI vs 1-week ISI) did not differ on any of the five aptitude covariates (musical aptitude, F (1, 67) < 0.001, p = .995; PM ability, i.e., SRT, F (1, 66) = 0.085, p =.772; WM, F (1, 67) = 1.155, p = .286; DM ability, i.e., CVMT, F (1, 66) = 0.036, p = .849; PSTM, i.e., NWR, F (1, 66) = 1.959, p = .166), suggesting the assumptions of independence of the covariates and treatment were met.

Table 30 presents a summary of results from the ANCOVA models conducted on TS3_Post_OPicN_4com_Acc. Model 7 included ISI and the musical aptitude construct score into modeling, and the result showed that musical aptitude was a significant predictor of outcome oral picture naming accuracy at the end of the last training session, F (1, 65) = 5.616, p = .021, η_p^2 = .080, when controlling for ISI. Model 8 further added the interaction between musical aptitude and ISI into modeling, and found that the interaction was not significant, F (1, 64) = 0.014, p = 0.906, η_p^2 < .001, suggesting that the effect of musical aptitude does not differ depending on the practice schedule (daily or weekly). The parameter estimates of

musical aptitude on TS3_Post_OPicN_4com_Acc based on Model 7 is presented in Table 31. The positive coefficient (B=0.034) for musical aptitude indicates that participants with higher musical aptitude performed better in oral picture naming accuracy by the end of training, suggesting that musical aptitude still plays a facilitative role in Mandarin oral word production even at this later stage by the end of training.

Model 9 tested the effects of PM ability as measured by SRT on oral picture naming accuracy performance at the end of training when controlling ISI; the results showed that PM ability was not a significant predictor, F (1, 65) = 2.341, p = .131, with small-to-medium effect size, η_p^2=.035. Model 10 further included the interaction term between SRT and ISI into modeling, and the interaction was not significant either, F (1, 64) = 0.108, p = 0.744, η_p^2= .002. Again, parameter estimates for SRT is presented in Table 31. These results suggest that PM ability did not turn out to predict oral Mandarin word production performance at this later stage by the end of the third training session.

In light of the finding that PM ability did not turn out to predict performance at the later stage by the end of the last training session, further analyses were conducted to check whether WM plays a role at this later stage. Model 11 included ISI and WM construct score into modeling, and found that WM was a significant predictor of TS3_Post_OPicN_4com_Acc, F (1, 65) = 5.827, p = .019, η_p^2 = .082, when controlling ISI. Model 12 further added the interaction term into modeling, and found that the WM-ISI interaction was not significant, F (1, 64) = 0.994, p = 0.323, η_p^2 = .015, indicating the effect of WM does not differ depending on ISI. The positive

coefficient for WM (*B*=0.031) (see Table 31) indicates that participants with higher WM capacity performed better in oral picture naming accuracy at the end of TS3, suggesting that WM capacity still plays a facilitative role in oral Mandarin word production at this later stage.

Table 30. *Summary of ANCOVA results on TS3_Post_OPicN_4com_Acc*

Model 7	TS3_Post_OPicN_4com_Acc		
df (1, 65)	*F*	*p*	η_p^2
ISI	0.000	0.995	0.000
Music_Z 2PitchSTM1Mus	5.616	0.021	0.080

Model 9	TS3_Post_OPicN_4com_Acc		
df (1, 65)	*F*	*p*	η_p^2
ISI	0.002	0.963	0.000
SRT_Z	2.341	0.131	0.035

Model 8	TS3_Post_OPicN_4com_Acc		
df (1, 64)	*F*	*p*	η_p^2
ISI	0.000	0.991	0.000
Music_Z 2PitchSTM1PTM	5.279	0.025	0.076
ISI * Music_Z 2PitchSTM1PTM	0.014	0.906	0.000

Model 10	TS3_Post_OPicN_4com_Acc		
df (1, 64)	*F*	*p*	η_p^2
ISI	0.001	0.973	0.000
SRT_Z	2.355	0.130	0.035
ISI * SRT_Z	0.108	0.744	0.002

Model 11	TS3_Post_OPicN_4com_Acc		
df (1, 65)	*F*	*p*	η_p^2
ISI	0.095	0.759	0.001
WM_Z ShapeBOspan	5.827	0.019	0.082

Model 12	TS3_Post_OPicN_4com_Acc		
df (1, 64)	*F*	*p*	η_p^2
ISI	0.110	0.741	0.002
WM_Z ShapeBOspan	6.420	0.014	0.091
ISI * WM_Z ShapeBOspan	0.994	0.323	0.015

Table 31. *Parameter Estimates of Aptitudes in Single-Covariate Analyses on TS3_Post_OPicN_4com_Acc with ISI controlled*

		TS3_Post_OPicN_4com_Acc				
Model	Variable	*B*	*SEB*	*t*	*p*	η_p^2
Model 7	Music_Z2PitchSTM1PTM	0.034	0.014	2.370	0.021	0.080
Model 9	SRT_Z	0.018	0.012	1.530	0.131	0.035
Model 11	WM_ZShapeBOspan	0.031	0.013	2.414	0.019	0.082

To test whether PM ability plays role in oral picture naming RT at this later stage (Hypothesis 2b-iii), when controlling for both ISI and individual differences in L1 word naming RT, Model 13 was conducted including ISI, English OWN RT and SRT into modeling. See Table 32 for a summary of results. L1/English oral word naming RT turned out to be a significant covariate for oral picture naming RT at the end of TS3, F (1,62) = 7.611, p = .008, η_p^2 = .109; however, SRT was not, F (1,62) = 2.077, p= .155, again with small-to-medium effect size, η_p^2 = .032. Additional analyses were conducted to check whether the two covariates interact with the experimental factor ISI, and the results showed that neither L1 word naming RT nor SRT interacted with ISI (see Models 14, and 15 in Table 32). Parameter estimates for SRT and L1 word naming RT on oral picture naming RT at the end of TS3 are presented in Table 33. The positive coefficient (B=0.759) for English word naming RT suggests that participants who were faster in L1 word naming were also faster in oral picture naming in L2 Mandarin Chinese at the end of TS3. PM ability, as measured by SRT, however, did not seem to play a role in oral picture naming RT in the post-session quiz in TS3.

Table 32. *Summary of ANCOVA results on TS3_Post_OPicN_RT_M_lg*

Model 13	TS3_Post_OPicN_RT_M_lg		
df (1, 62)	*F*	*p*	η_p^2
ISI	0.255	0.616	0.004
EngOWN_RT_M_lg	7.611	0.008	0.109
SRT_Z	2.077	0.155	0.032

Model 14	TS3_Post_OPicN_RT_M_lg		
df (1, 61)	*F*	*p*	η_p^2
ISI	0.096	0.757	0.002
EngOWN_RT_M_lg	6.596	0.013	0.098
SRT_Z	2.138	0.149	0.034
ISI * EngOWN_RT_M_lg	0.103	0.750	0.002

Model 15	TS3_Post_OPicN_RT_M_lg		
df (1, 61)	*F*	*p*	η_p^2
ISI	0.260	0.612	0.004
EngOWN_RT_M_lg	8.500	0.005	0.122
SRT_Z	1.888	0.174	0.030
ISI * SRT_Z	1.360	0.248	0.022

Table 33. *Parameter Estimates for PM and L1 WN RT* on *TS3_Post_OPicN_RT_M when controlling for ISI*

		TS3_Post_OPicN_RT_M_lg				
Model	Variable	*B*	*SEB*	*t*	*p*	η_p^2
Model 13	EngOWN_RT_M_lg	0.759	0.275	2.759	0.008	0.109
	SRT_Z	0.024	0.016	1.441	0.155	0.032

6.5 Effects of ISI and RI on Retention

The third research question concerns the effects of ISI and RI on L2 *retention* of Mandarin tonal word production; this section thus focuses on reporting results concerning outcome performance on the retention test administered on Day 5. Four sub- research questions were asked: RQ3-a concerns the effects of ISI and RI on the

retention of the declarative component of Mandarin word production, i.e., picture-Pinyin associations, as measured by written picture naming accuracy; RQ3-b concerns the effects of ISI and RI on the retention of oral tone production skill, from Pinyin to articulation, on old words participants had practiced throughout the training sessions, as measured by oral word naming accuracy and RT on old words; RQ3-c regards the effects of ISI and RI on the retention of oral tone production skill on new words (i.e., words they never practiced before), as measured by oral word naming accuracy and RT on new words; and lastly RQ3-d concerns the effects of ISI and RI on the retention of oral Mandarin word production, from meaning to articulation, as measured by oral picture naming accuracy and RT.

To test the effects of ISI and RI on each of the seven outcome measures from the retention test, a set of two-way ANOVAs, with ISI and RI as two independent variables, were first conducted on each of the seven outcome measures, to compare group differences without taking into consideration of individual differences in aptitudes. A second step was further taken to conduct a set of ANCOVAs, again with ISI and RI as two independent variables, plus the set of five aptitude construct scores as covariates, in order to see when controlling individual differences in aptitudes, whether the patterns regarding the effects of ISI and RI change. When the outcome or dependent variable was an RT measure, in addition to the five aptitude construct scores, L1 word naming RT was also included into modeling and therefore controlled. Results in response to each of the four sub-RQs are presented in order.

6.5.1 Retention of declarative word knowledge (as measured by written picture naming)

It was hypothesized that the effects of temporal distribution of practice on the retention of declarative knowledge may be determined by the optimal ISI/RI ratios. As Group A's and Group D's ISI/RI ratios (14% and 25%, respectively) fall within the optimal range (10%-30%) while Group B's and Group C's fall outside of the optimal range, Group A (ISI-1d; RI-1w) and Group D (ISI-1w; RI-4w) were expected to outperform Group B (ISI-1d; RI-4w) and Group C (ISI-1w; RI-1w) on written picture naming accuracy (the measure of retention of declarative knowledge) administered on Day 5. In other words, an interaction between ISI and RI were expected on the retention of declarative knowledge as measured by written picture naming accuracy.

Results from the ANOVA and ANCOVA with written picture naming accuracy as the dependent variable are presented in Table 34. Model 1 tests the effects of ISI and RI without controlling aptitudes, and Model 2 tests the effects of ISI and RI while controlling for individual differences in aptitudes. In Model 1, the Levene's test showed that the equality of variances assumption was not met (p = .036); however, the largest SD (.17) was less than two times the smallest SD (.11), and therefore, ANOVA was considered robust. Model 1 yielded a significant main effect for RI, F (1, 64) = 22.779, $p < .001$, η_p^2=.262, a marginally significant main effect for ISI, F (1, 64) = 3.951, p = .051, η_p^2=.058, and a nonsignificant interaction between ISI and RI, F (1, 64) = 1.838, p = .180, η_p^2=.028. In Model 2, when individual differences in aptitudes were controlled, the main effect for RI was still

significant, F (1, 59) = 25.011, $p < .001$, with even larger effect size, η_p^2=.298, while the main effect of ISI turned out to be nonsignificant, F (1, 59) = 1.694, p = .198, η_p^2=.028. The interaction between ISI and RI in Model 2 was not significant either, F (1, 59) = 1.262, p = .266, η_p^2=.021. Figure 9 presents the estimated marginal means adjusted for the covariates from Model 2. It shows that the 4-week RI resulted in much lower performance in written picture naming accuracy than the 1-week RI, i.e., Groups A and C outperformed Groups B and D.

Table 34. *Effects of ISI and RI on retention of declarative word knowledge as measured by written picture naming accuracy*

Model 1	D5_WPicN_4com_Acc		
df (1, 64)	*F*	*p*	η_p^2
ISI	3.951	0.051	0.058
RI	22.779	0.000	0.262
ISI*RI	1.838	0.180	0.028

Model 2	D5_WPicN_4com_Acc		
df (1, 59)	*F*	*p*	η_p^2
ISI	1.694	0.198	0.028
RI	25.011	0.000	0.298
ISI * RI	1.262	0.266	0.021
CVMT_ZAcqDelayed	1.029	0.315	0.017
WM_ZShapeBOspan	8.775	0.004	0.129
NWR_Z	2.537	0.117	0.041
Music_Z2PitchSTM1PTM	0.937	0.337	0.016
SRT_Z	0.076	0.784	0.001

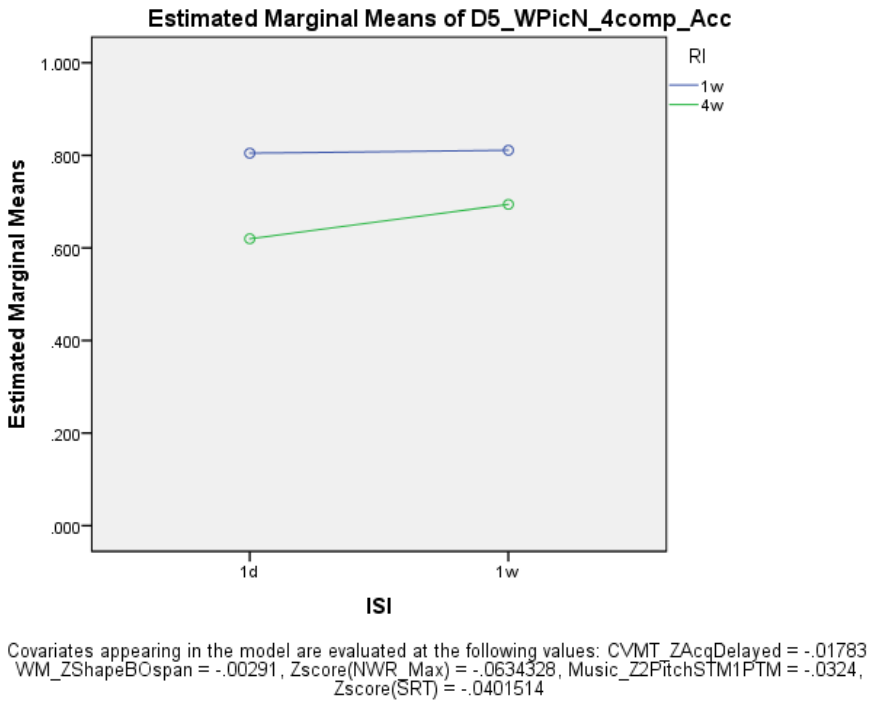


Figure 9. Estimated marginal means of D5_WPicN_4com_Acc

The results disconfirmed the hypothesis for an interaction between ISI and RI (Hypothesis 3a) on the retention of declarative knowledge; it was not that 1-day ISI combined with 1-week RI group (Group A) and 1-week ISI combined with 4-week RI group (Group D) which have the optimal ISI/RI ratios outperformed the other two groups with suboptimal ISI/RI ratios. Instead, the main effect of RI was robust, with very large effect size, η_p^2=.298, showing convincingly that longer retention interval will result in more forgetting of declarative knowledge. The main effect of ISI was nonsignificant when controlling for individual differences in aptitudes, suggesting that temporal distribution of practice (daily vs. weekly) did not exert a big role for 1-week or 4-week retention of the declarative word knowledge. Descriptively (see Figure 9), it seems that ISI (1-day or 1-week) plays little role for the short-term 1-week retention; for the long-term 4-week retention, however, spacing (weekly vs

daily) seems to facilitate retention of declarative knowledge (note however that the interaction was not significant).

6.5.2 Retention of tone production skill (as in oral word naming) on old words

Moving on to the effects of ISI and RI on the retention of a skill, i.e., tone production on old words, in contrast to the retention of declarative knowledge, it was hypothesized that the massed practice schedule, which is more likely to enable participants to move from the second to the third stage of skill acquisition, should work better than the distributed practice schedule, irrespective of retention intervals since the developed procedural knowledge was considered robust to memory decay. It was expected the 1-day ISI groups (Groups A and B) would outperform the 1-week ISI groups (Groups C and D) in oral word naming accuracy on old words (Hypothesis 3b-i). As for RT performance, similarly, the 1-day ISI groups were expected to respond faster than the 1-week ISI group on the retention test (Hypothesis 3b-ii). That is, a main effect of ISI was expected on both the accuracy and the RT measures in oral word naming on old words.

Results from the ANOVAs and ANCOVAs testing the effects of ISI and RI on the retention of tone production skill on old words in terms of accuracy and RT respectively are presented in Table 35. Model 3 tests the effects of ISI and RI on tone production accuracy on old words in oral word naming (D5_OWN_old_Tone_Acc), without controlling aptitudes, and Model 4 tests the effects of ISI and RI on the same outcome while controlling for individual difference in aptitudes. In both models, the homogeneity of variances assumption was met, according to Levene's test (p=.111 in Model 3, and p=.606 in Model 4). The results from Model 3 showed that the main

effect of ISI was not significant, F (1, 64) = 0.307, p = .582, η_p^2=.005; the main effect of RI was not significant either, F (1, 64) = 0.531, p = .469, η_p^2=.008. The interaction between ISI and RI was also nonsignificant, F (1, 64) = 0.117, p = .733, η_p^2=.001. The results from Model 4 showed that, after controlling for individual differences in aptitudes, the pattern of the effects did not change; Model 4 yielded a nonsignificant main effect for ISI, F (1, 59) = 1.418, p = .238, η_p^2=.023, a nonsignificant main effect for RI, F (1, 59) = 0.565, p = .455, η_p^2=.009, and a nonsignificant interaction between ISI and RI, F (1, 59) = 0.266, p = .608, η_p^2=.004. Figure 10 presents the estimated marginal means of tone production accuracy in oral word naming on practiced words; the four groups seem to perform at the same level, irrespective of RI or ISI.

Table 35. *Effects of ISI and RI on retention of tone production skill on old words*

Model 3	D5_OWN_old_Tone_Acc		
df (1, 64)	*F*	*p*	η_p^2
ISI	0.307	0.582	0.005
RI	0.531	0.469	0.008
ISI * RI	0.117	0.733	0.002

Model 5	D5_OWN_old_RT_M_lg		
df (1, 60)	*F*	*p*	η_p^2
ISI	1.518	0.223	0.025
RI	0.016	0.898	0.000
ISI * RI	1.411	0.240	0.023

Model 4	D5_OWN_old_Tone_Acc		
df (1, 59)	*F*	*p*	η_p^2
ISI	1.418	0.238	0.023
RI	0.565	0.455	0.009
ISI * RI	0.266	0.608	0.004
CVMT_Z	0.001	0.979	0.000
AcqDelayed WM_Z	2.962	0.090	0.048
ShapeBOspan NWR_Z	3.347	0.072	0.054
Music_Z	4.839	0.032	0.076
2PitchSTM 1PTM SRT_Z	1.594	0.212	0.026

Model 6	D5_OWN_old_RT_M_lg		
df (1, 54)	*F*	*p*	η_p^2
ISI	1.078	0.304	0.020
RI	0.103	0.750	0.002
ISI * RI	0.744	0.392	0.014
CVMT_Z	0.367	0.547	0.007
AcqDelayed WM_Z	0.140	0.710	0.003
ShapeBOspan NWR_Z	0.217	0.643	0.004
Music_Z	0.206	0.651	0.004
2PitchSTM 1PTM SRT_Z	0.397	0.531	0.007
EngOWN_RT _M_lg	12.207	0.001	0.184

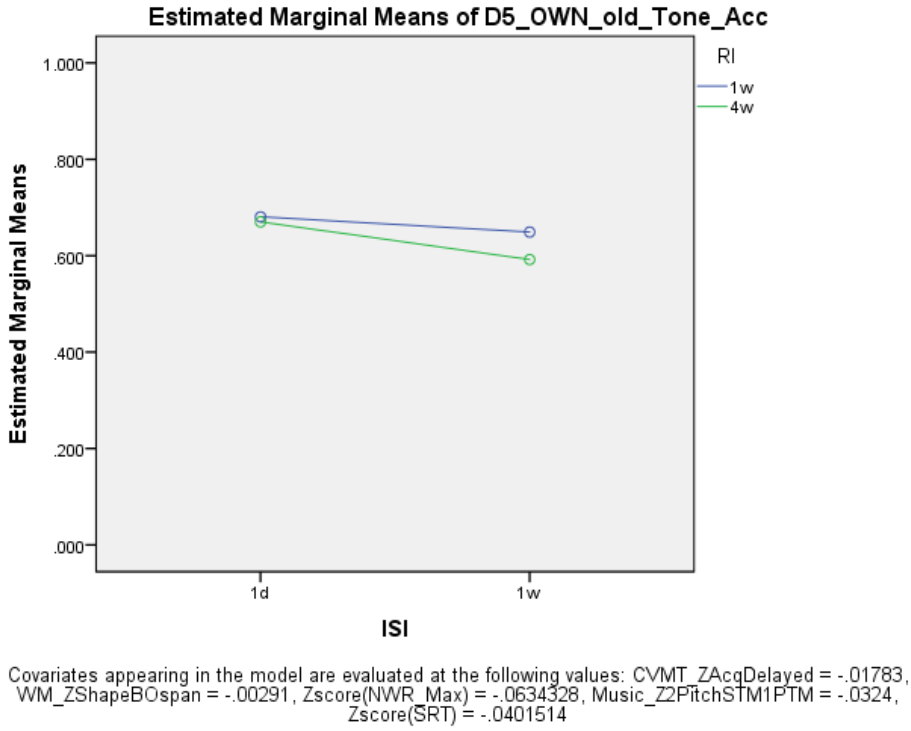


Figure 10. Estimated marginal means of D5_OWN_old_Tone_Acc

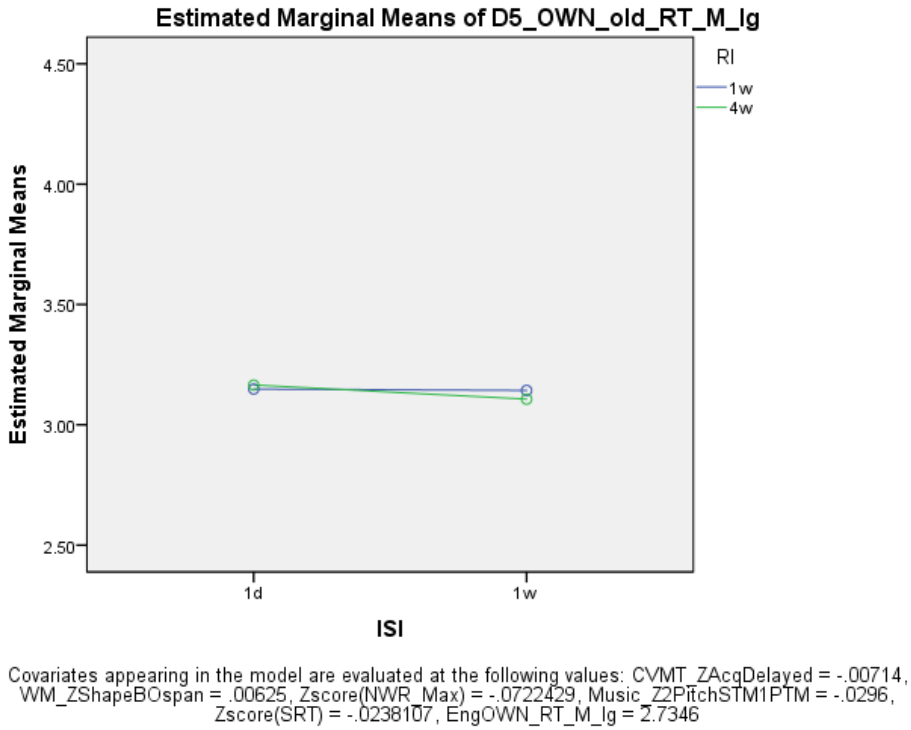


Figure 11. Estimated marginal means of D5_OWN_old_RT_M_lg

With regard to RT performance, Model 5 tests the effects of ISI and RI on oral word naming RTs (log transformed) on old words (D5_OWN_old_RT_M_lg), without controlling aptitudes, and Model 6 tests the effects of ISI and RI on the same outcome while controlling for individual difference in aptitudes and L1 word naming RT. In both models, the homogeneity of variances assumption was met, according to Levene's test (p = .699 in Model 5, p = .176 in Model 6). Model 5 yielded a nonsignificant main effect for ISI, F (1, 60) = 1.518, p = .223, η_p^2= .025, a nonsignificant main effect for RI, F (1, 60) = 0.016, p = .898, η_p^2< .001, and a nonsignificant interaction between ISI and RI, F (1, 60) = 1.411, p = .240, η_p^2= .023. The results from Model 6 showed that, after controlling for individual differences in aptitudes and L1 word naming RT, the pattern for the effects of ISI and RI did not change, i.e., again, the main effect of ISI was nonsignificant, F (1, 54) = 1.078, p = .304, η_p^2= .020, the main effect of RI was nonsignificant, F (1, 54) = 0.103, p = .750, η_p^2= .002, and the interaction between ISI and RI was nonsignificant either, F (1, 54) = 0.744, p = .392, η_p^2= .014. Figure 11 presents the estimated marginal means of oral word naming RT on practiced words; the four groups seem to perform at the same level in RT, irrespective of RI or ISI.

While a main effect of ISI was expected on both the accuracy and the RT measures in oral word naming on old words, these hypotheses did not borne out. Instead, the main effect of ISI was not observed, neither was there a main effect of RI or an interaction between ISI and RI. In fact, the four groups performed at the same level in terms of both oral tone production accuracy and oral word naming RT on practiced words.

6.5.3 Retention of tone production skill (as in oral word naming) on new words

With regard to the effects of ISI and RI on the retention of tone production skill on new words, as no previous research has looked at these effects on the generalization of practiced skills in new contexts, no strong hypothesis was formulated. In the context of this study, as orally naming new Mandarin words from Pinyin involves fusing tone production with new combinations of segments, declarative knowledge about how to pronounce the tones is likely to be involved in this fusing process. As this task involves the retention of both declarative knowledge about how to pronounce tones and procedural knowledge in articulation, due to these double constraints (declarative knowledge was expected to be better retained in Groups A and D than Groups B and C, and procedural knowledge better retained in Groups A and B than Groups C and D), Group A might outperform the other three groups in oral word naming accuracy on new words, at least descriptively. No predictions were made about the effect of RI, ISI, or their interaction, on oral word naming RT on new words.

To explore the effects of ISI and RI on tone production skill in new words, an ANOVA was first conducted with ISI and RI as independent variables, and tone production accuracy in oral word naming on new words as dependent variable (Model 7). An ANCOVA was then conducted to see whether the effects of ISI and RI would change after controlling for individual differences in aptitudes by further including the five aptitude construct scores into modeling (Model 8). Similar analyses were conducted on the RT outcome measure from oral word naming on new words, with Model 9 only including the main effects of ISI and RI and the interaction between

them and Model 10 further adding the five aptitude construct scores and L1 word naming RT. In all these models, the homogeneity of variances assumption was met in each model, according to Levene's test (p=.225 in Model 7, p=.446 in Model 8, p=.390 in Model 9, and p= .242 in Model 10). The F-test results from these four models are presented in Table 36.

Model 7 tests the effects of ISI and RI on tone production accuracy on new words in oral word naming (D5_OWN_new_Tone_Acc), without controlling aptitudes. The results showed a nonsignificant main effect for ISI, F (1, 64) = 0.657, p = .421, η_p^2= .010, a nonsignificant main effect for RI, F (1, 64) = 2.910, p = .093, η_p^2= .043, and a nonsignificant interaction between ISI and RI, F (1, 64) = 0.126, p = .724, η_p^2= .002. In Model 8, when controlling for individual differences in aptitudes, however, the main effect for ISI turned out to be significant, F (1, 59) = 4.106, p =.047, η_p^2=.065, though the main effect for RI remained marginal, F (1, 59) = 3.606, p =.062, η_p^2=.058. The interaction between ISI and RI remained nonsignificant, F (1, 59) = 0.453, p =.504, η_p^2=.008. The reason why the main effect for ISI only appeared when controlling for individual differences in aptitudes was because partialing out these individual differences reduced the unexplained variance and therefore increased the power to detect the treatment effect. Figure 12 presents the marginal means of tone production accuracy in oral word naming on new/generalization words, after adjusted for the covariates. The significant main effect for ISI suggests that the daily practice schedule led to significantly better tone production accuracy in oral word naming on new words, after controlling for individual differences in aptitudes. In other words, the advantage of massed practice showed up on tone production

accuracy on NEW words, though such effect was hypothesized but not borne out on tone production accuracy on old words. Descriptively, from Figure 12, there seems to be a main effect for RI, as 4-week RI seems to have resulted in lower tone production accuracy than 1-week RI at each of the respective ISI levels.

Table 36. *Effects of ISI and RI on retention tone production skill in new words*

Model 7	D5_OWN_new_Tone_Acc		
df (1, 64)	*F*	*p*	η_p^2
ISI	0.657	0.421	0.010
RI	2.910	0.093	0.043
ISI * RI	0.126	0.724	0.002

Model 8	D5_OWN_new_Tone_Acc		
df (1, 59)	*F*	*p*	η_p^2
ISI	4.106	0.047	0.065
RI	3.606	0.062	0.058
ISI * RI	0.453	0.504	0.008
CVMT_Z	0.006	0.941	0.000
AcqDelayed WM_Z	15.267	0.000	0.206
ShapeBOspan NWR_Z	6.851	0.011	0.104
Music_Z	7.102	0.010	0.107
2PitchSTM 1PTM SRT_Z	3.622	0.062	0.058

Model 9	D5_OWN_new_RT_M_lg		
df (1, 63)	*F*	*P*	η_p^2
ISI	0.335	0.565	0.005
RI	0.000	0.984	0.000
ISI * RI	0.514	0.476	0.008

Model 10	D5_OWN_new_RT_M_lg		
df (1, 57)	*F*	*P*	η_p^2
ISI	0.778	0.381	0.013
RI	0.027	0.870	0.000
ISI * RI	0.596	0.443	0.010
CVMT_Z	0.188	0.666	0.003
AcqDelayed WM_Z	0.079	0.780	0.001
ShapeBOspan NWR_Z	3.755	0.058	0.062
Music_Z	1.160	0.286	0.020
2PitchSTM 1PTM SRT_Z	1.415	0.239	0.024
EngOWN_RT_M_lg	2.314	0.134	0.039

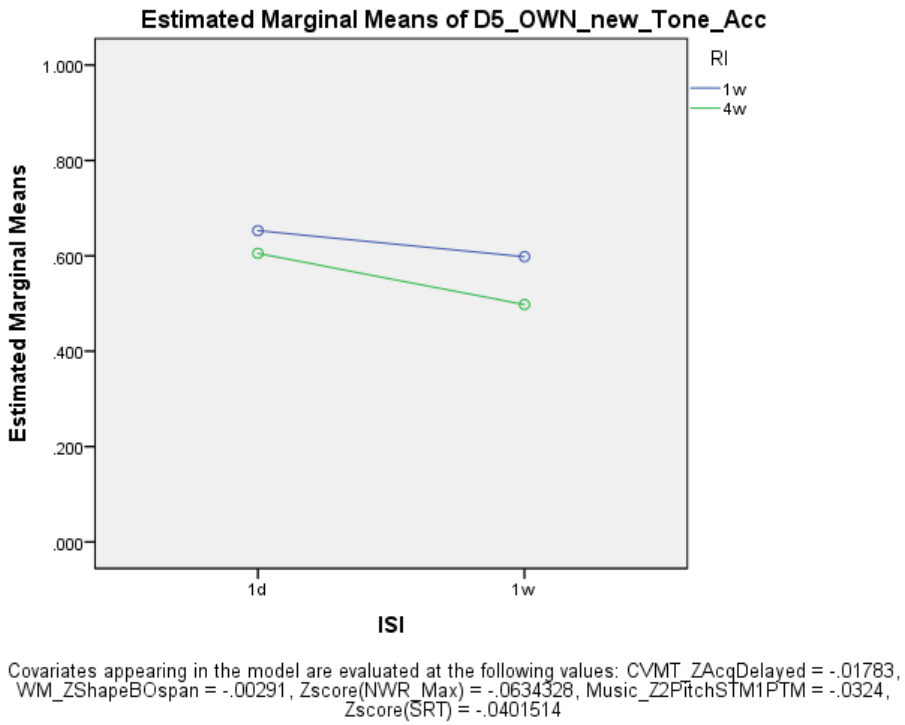


Figure 12. Estimated marginal means of D5_OWN_new_Tone_Acc

Model 9 tests the effects of ISI and RI on oral word naming RT on new words, without controlling individual differences. The ANOVA yielded a nonsignificant main effect for ISI, F (1, 63) = 0.335, p = .565, η_p^2= .005, a nonsignificant main effect for RI, F (1, 63) < 0.001, p = .984, η_p^2<.001, and a nonsignificant interaction between ISI and RI, F (1, 63) = 0.514, p = .476, η_p^2= .008. When controlling for individual differences in aptitudes and L1 word naming RT, Model 10 also yielded a nonsignificant main effect for ISI, F (1, 57) = 0.778, p = .381, η_p^2= .013, a nonsignificant main effect for RI, F (1, 57) = 0.027, p = .870, η_p^2< .001, and a nonsignificant interaction between ISI and RI, F (1, 57) = 0.596, p = .443, η_p^2= .010. Figure 13 presents the estimated marginal means of oral word naming RT on new words after adjusting for the covariates. The four groups seem to be at the same level in RT, irrespective of ISI or RI.

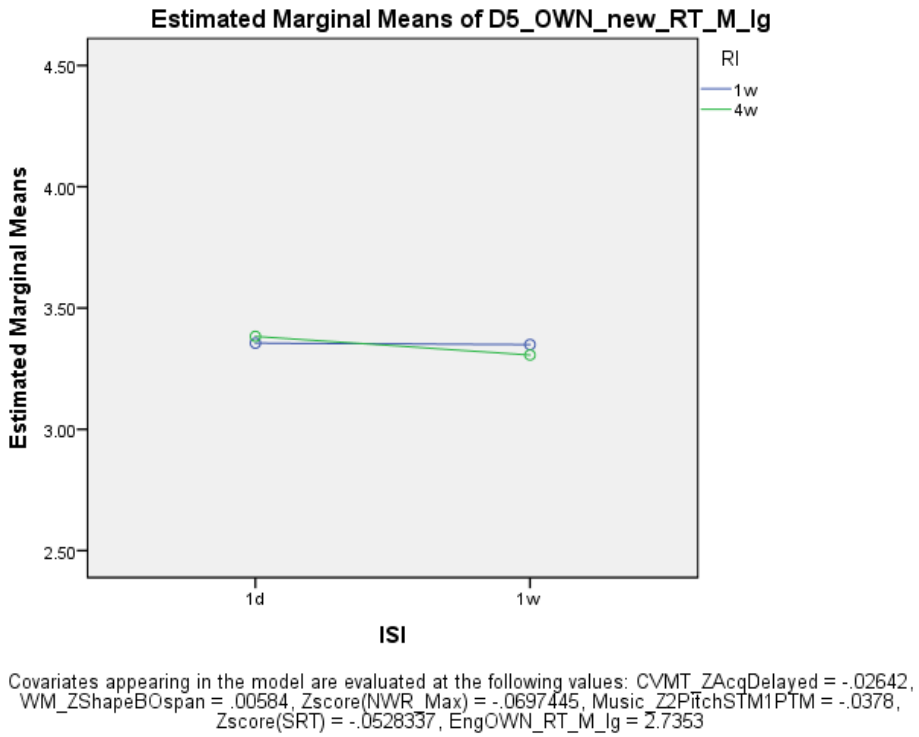


Figure 13. Estimated marginal means of D5_OWN_new_RT_M_lg

These analyses on the effects of ISI and RI on oral word naming in new words were exploratory. For oral tone production accuracy on new words, Group A performed better than the other three groups descriptively, as anticipated. In addition, a significant main effect for ISI was found on oral tone production accuracy on new words after controlling for individual differences in aptitudes, suggesting an advantage of massed practice over distributed practice on the retention of tone production skill (accuracy). As for the RT measure, the four groups seemed to perform at the same level, with neither ISI nor RI playing a role.

6.5.4 Retention of oral word production skill (as in oral picture naming)

With regard to the effects of ISI and RI on the retention of the oral word production skill from conceptualization to articulation, which is measured by the oral

picture naming task on Day 5, no strong hypothesis was formulated. In the context of this study, I anticipated that performance on oral picture naming may be first constrained by the retention of the declarative component of word knowledge (i.e., picture-Pinyin mappings) and then by the retention of the procedural component in oral production. Due to these double constraints (declarative knowledge was expected to be better retained in Groups A and D than Groups B and C, and procedural knowledge better retained in Groups A and B than Groups C and D), Group A may outperform the other three groups in oral picture naming accuracy, at least descriptively. Among the other three groups, Group D might outperform the Groups B and C in oral picture naming accuracy (because Group D was expected to have better retention of declarative word knowledge which is the prerequisite for correct production from meaning to form). No predictions were made about the effect of RI, ISI, or their interaction, on oral word naming RT on new words.

To explore the effects of ISI and RI on oral picture naming performance, separate ANOVAs were conducted on the accuracy and RT measures in oral picture naming completed on Day 5, with ISI and RI as independent variables. Additional separate ANCOVAs were conducted on the accuracy and RT measures of oral picture naming, to see whether the effects of ISI and RI would change when controlling for individual differences by further adding the five aptitude construct scores into modeling, plus L1 word naming RT when the outcome was the RT measure. Table 37 presents the *F*-test results from the four models. In Model 11 and Model 12, when the outcome measure was oral picture naming accuracy, the homogeneity of variances assumption was not met, according to Levene's test (p = .034 for Model 11, p = .042

for Model 12). However, as the largest SD (.16) was two times less than the smallest SD (.09), the ANOVA or ANCOVA was considered robust. In Model 13 and Model 14, when the outcome measure was oral picture naming RT, the homogeneity of variances assumption was met, according to Levene's test (p = .577 in Model 13, p = .557 in Model 14).

Model 11 tests the effects of ISI and RI on oral picture naming accuracy (D5_OPicN_4com_Acc), without controlling individual differences in aptitudes. The results showed a significant main effect for RI, F (1, 64) = 15.562, p < .001, η_p^2= .196, a nonsignificant main effect for ISI, F (1, 64) = 0.611, p = .437, η_p^2= .009, and a nonsignificant interaction between ISI and RI, F (1, 64) = 2.664, p = .108, η_p^2= .040. In Model 12, when controlling for individual differences in aptitudes, the pattern of results remained the same; that is, the main effect of RI was significant, F (1, 59) = 17.160, p < .001, with even larger effect size, η_p^2= .225, the main effect of ISI was nonsignificant, F (1, 59) = 0.083, p = .774, η_p^2= .001, and the interaction between ISI and RI remained nonsignificant, F (1, 59) = 2.535, p = .117, η_p^2= .041. Figure 14 presents the estimated marginal means of oral picture naming accuracy adjusted for the covariates. The pattern of the results showed that longer 4-week RI resulted in much lower performance in oral picture naming accuracy than the shorter 1-week RI when controlling individual differences in aptitudes.

When it comes to the RT performance, Model 13 tests the effects of ISI and RI without controlling for individual differences. The ANOVA yielded a significant main effect for ISI, F (1, 57) = 9.529, p = .003, η_p^2= .143, and for RI, F (1, 57) = 4.471, p = .039, η_p^2= .073, and a nonsignificant interaction between ISI and RI, F (1,

57) = 1.274, p = .264, η_p^2= .022. When controlling for individual differences in aptitudes and L1 word naming RT, Model 14 yielded the same pattern of results, i.e., a significant main effect for ISI, F (1, 51) = 10.287, p = .002, η_p^2= .168, and for RI, F (1, 51) = 4.479, p = .039, η_p^2= .081, and a nonsignificant interaction between ISI and RI, F (1, 51) = 0.538, p = .467, η_p^2= .010. Figure 15 presents the estimated marginal means of oral picture naming RT after adjusted for the covariates. It seems that the longer 4-week RI resulted in much slower RT than the shorter 1-week RI, irrespective of ISI, suggesting that longer RI resulted in more delay in oral picture naming automaticity. In contrast, the longer 1-week ISI resulted in much faster RT than the shorter 1-day ISI, irrespective of RI; in other words, distributed practice resulted in faster RT than massed practice on retention performance in oral picture naming.

Table 37. *Effects of ISI and RI on retention of oral word production skill as in oral picture naming*

Model 11	D5_OPicN_4com_Acc		
df (1, 64)	*F*	*p*	η_p^2
ISI	0.611	0.437	0.009
RI	15.562	0.000	0.196
ISI * RI	2.664	0.108	0.040

Model 13	D5_OPicN_RT_M_lg		
df (1, 57)	*F*	*p*	η_p^2
ISI	9.529	0.003	0.143
RI	4.471	0.039	0.073
ISI * RI	1.274	0.264	0.022

Model 12	D5_OPicN_4com_Acc		
df (1, 59)	*F*	*p*	η_p^2
ISI	0.083	0.774	0.001
RI	17.160	0.000	0.225
ISI * RI	2.535	0.117	0.041
CVMT_Z	0.536	0.467	0.009
AcqDelayed			
WM_Z	4.147	0.046	0.066
ShapeBOspan			
NWR_Z	0.794	0.376	0.013
Music_Z	4.710	0.034	0.074
2PitchSTM			

Model 14	D5_OPicN_RT_M_lg		
df (1, 51)	*F*	*p*	η_p^2
ISI	10.287	0.002	0.168
RI	4.479	0.039	0.081
ISI * RI	0.538	0.467	0.010
CVMT_Z	0.329	0.569	0.006
AcqDelayed			
WM_Z	0.001	0.976	0.000
ShapeBOspan			
NWR_Z	1.054	0.310	0.020
Music_Z	0.003	0.956	0.000
2PitchSTM			

1PTM SRT_Z	0.603	0.441	0.010	1PTM SRT_Z	1.835	0.182	0.035
				EngOWN_RT_M_lg	8.89	0.004	0.148

These analyses on the effects of ISI and RI on retention performance in oral picture naming were exploratory. Descriptively, it was indeed the case that Group A outperformed the other three groups in oral picture naming accuracy, as anticipated. However, Group D did not outperform Groups B and C in oral picture naming accuracy; instead, Group C, the group with short-term 1-week RI outperformed Groups B and D, the groups with longer-term 4-week RI, descriptively in terms of group means. As for retention performance on the RT measure, longer RI resulted in slower RT, but longer ISI resulted in faster RT. These results will be discussed in the next section.

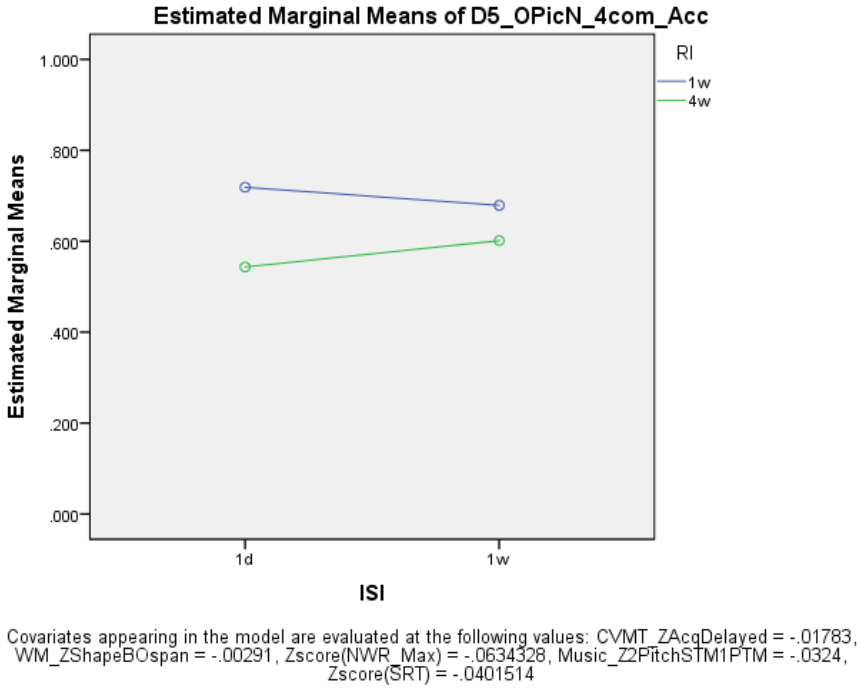


Figure 14. Estimated marginal means of D5_OPicN_4com_Acc

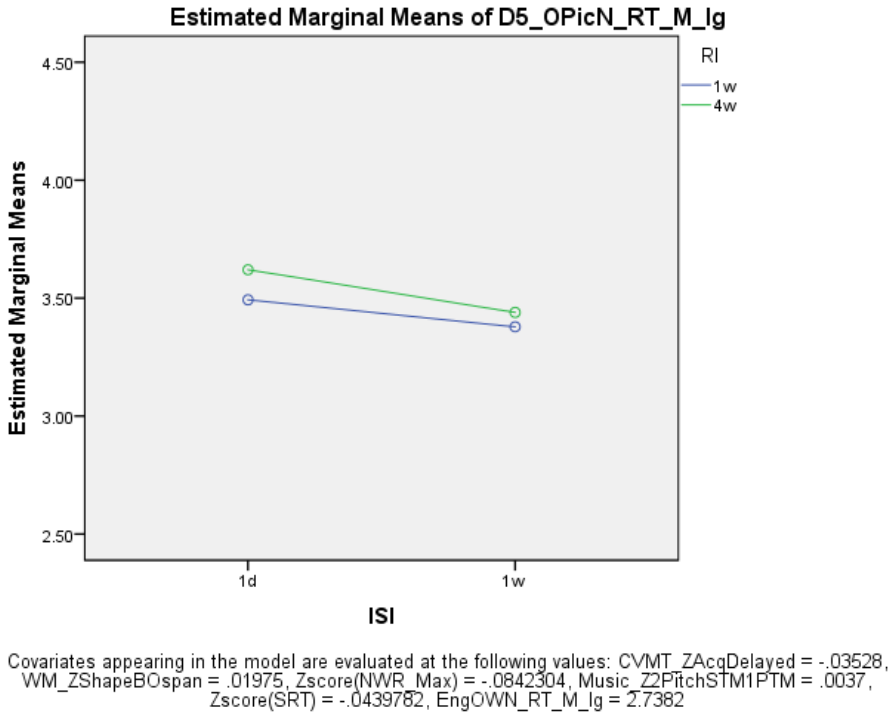


Figure 15. Estimated marginal means of D5_OPicN_RT_M_lg

6.6 Effects of Aptitudes on Retention

The fourth research question concerns the effects of cognitive aptitudes on L2 retention of Mandarin tonal word production when controlling for the effects of ISI and RI. Four sub-questions were asked: RQ4-a regards the role of cognitive aptitudes on the retention of declarative word knowledge (i.e., picture-Pinyin mappings); RQ4-b concerns the role of aptitudes on the retention of tone production skill in practice words; RQ4-c concerns the role of aptitudes on the retention of tone production skill in new words; and finally RQ4-d concerns the role of cognitive aptitudes on the retention of oral word production skill as in oral picture naming.

Specific hypotheses were put forward for each of the sub-questions (see Chapter 4). Each of the hypotheses was tested by conducting ANCOVA models in two steps. In order to control any effects of ISI and RI on outcome performance, the main effects of ISI and RI and the interaction term between ISI and RI were kept in all modeling. When the outcome measure was an RT measure, individual differences in L1 word naming RT were further controlled by including that variable into modeling. To test each hypothesis, an ANCOVA was first run with the hypothesized aptitude construct score as a covariate (without interaction) in addition to the control variables, to test how the aptitude covariate is related to the outcome measure. The next step was to conduct another ANCOVA with the same set of control variables and the covariate, plus allowing for interactions between the covariate and the experimental factors (i.e., ISI and RI) to check whether there was an ATI effect, i.e., whether the effect of the aptitude on outcome differed depending on the experimental conditions. Recall that the four experimental groups did not differ on any of the five aptitude covariates (see Table 13 under Section 6.1), suggesting that the assumption of independence of the covariates and treatment was met. Levene's test was conducted for each ANCOVA model; according to this test, homogeneity of variances assumption was met in some models but not met in some other models. For the models that the homogeneity assumption was not met according to Levene's test, the SDs of the groups on the outcome measure was checked; in none of the models was the largest SD more than two times of the smallest SD. Therefore, the ANCOVAs were considered robust.

Appendix N presents the correlations between the five aptitude construct scores and the outcome measures on the retention test in terms of both accuracy and RT across the four experimental groups. The ANCOVA results in response to each of the four sub-RQs are presented in order.

6.6.1 Retention of declarative word knowledge (as measured by written picture naming)

Hypothesis 4a states that DM ability plays a facilitative role in the retention of declarative word knowledge (i.e., picture-Pinyin mappings) as measured by written picture naming accuracy. To test this hypothesis, Model 1 was conducted with ISI and RI (and their interaction) as controlling variables, the DM ability construct score (CVMT_ZAcqDelayed) as the covariate, and D5_WPicN_4com_Acc as the dependent variable. The results showed that CVMT was significantly related to the retention performance in written picture naming accuracy, F (1, 63) = 4.198, p = .045, η_p^2 = .062, after controlling the effects of ISI and RI. Model 2 was conducted to check whether DM ability interacted with the experimental conditions, and the results showed that CVMT did not interact with ISI, F (1, 61) = 0.026, p = .873, η_p^2 < .001, nor with RI, F (1, 61) = 2.423, p = .125, η_p^2 = .038, suggesting that the relationship between CVMT and written picture naming accuracy on Day 5 did not differ at different levels of ISI or RI. A summary of the results from the models is presented in Table 38. The parameter estimates of CVMT on D5_WPicN_4com_Acc from Model 1 are presented in Table 39. Parameter estimates from Model 1 were retained rather than from Model 2 because the interactions tested in Model 2 were not statistically significant. The positive coefficient of CVMT on D5_WPicN_4com_Acc (B = 0.038)

suggest that participants with higher DM ability achieved better retention performance in written picture naming accuracy. This finding confirmed Hypothesis 4a that DM ability plays a facilitative role on the retention of declarative word knowledge (i.e., picture-Pinyin mappings) as measured by written picture naming accuracy.

Though I did not make a prior hypothesis regarding the role of WM capacity on retention performance in written picture naming accuracy, a follow-up analysis was conducted to see whether WM also plays a role in the retention of declarative word knowledge in terms of meaning-spelling mappings. See Model 3 and Model 4 in Table 38. The results showed that WM was significantly related to written picture naming accuracy on the retention test, F (1, 63) = 9.681, p = .003, η_p^2 = .133, and WM did not interact with either ISI or RI. See the parameter estimates in Table 39. The positive coefficient of WM on D5_WPicN_4com_Acc (B = 0.056) suggests that WM capacity plays a facilitative role in the retention of declarative word naming.

The above results show that both DM ability and WM capacity play a facilitative role in the retention of declarative word knowledge, in isolation. In order to see whether DM ability still plays a role when WM is controlled, Model 5 was conducted by including both DM and WM into modeling at the same time (see Table 38). The results showed that conditioned on WM, CVMT turned out to be a nonsignificant covariate, F (1, 62) = 1.377, p = .245, η_p^2 = .022. Instead, WM was still a robust predictor, F (1, 62) = 6.548, p = .013, η_p^2 = .096, when controlling for CVMT. The parameter estimates of the two covariates from Model 5 are presented in

Table 40. WM capacity plays a facitative role independently from DM ability on the retention of declarative knowledge.

Table 38. *Summary of ANCOVA models testing the effects of aptitudes on D5_WPicN_4com_Acc when controlling IS and RI*

Model 1	D5_WPicN_4com_Acc		
df (1, 63)	*F*	*p*	η_p^2
ISI	3.952	0.051	0.059
RI	25.698	0.000	0.290
ISI * RI	1.860	0.178	0.029
CVMT_Z AcqDelayed	4.198	0.045	0.062

Model 3	D5_WPicN_4com_Acc		
df (1, 63)	*F*	*p*	η_p^2
ISI	2.861	0.096	0.043
RI	24.455	0.000	0.280
ISI * RI	1.856	0.178	0.029
WM_Z ShapeBOspan	9.681	0.003	0.133

Model 2	D5_WPicN_4com_Acc		
df (1, 61)	*F*	*p*	η_p^2
ISI	4.037	0.049	0.062
RI	26.598	0.000	0.304
ISI * RI	1.899	0.173	0.030
CVMT_Z AcqDelayed	5.665	0.020	0.085
ISI * CVMT_Z AcqDelayed	0.026	0.873	0.000
RI * CVMT_Z AcqDelayed	2.423	0.125	0.038

Model 4	D5_WPicN_4com_Acc		
df (1, 61)	*F*	*p*	η_p^2
ISI	2.735	0.103	0.043
RI	23.887	0.000	0.281
ISI * RI	1.506	0.224	0.024
WM_Z ShapeBOspan	9.218	0.004	0.131
ISI * WM_Z ShapeBOspan	0.125	0.725	0.002
RI * WM_Z ShapeBOspan	0.478	0.492	0.008

Modle 5	D5_WPicN_4com_Acc		
df (1, 62)	*F*	*p*	η_p^2
ISI	2.961	0.090	0.046
RI	25.653	0.000	0.293
ISI * RI	1.855	0.178	0.029
CVMT_Z AcqDelayed	1.377	0.245	0.022
WM_Z ShapeBOspan	6.548	0.013	0.096

Table 39. *Parameter Estimates of Aptitudes in Single-Covariate Analyses on D5_WPicN_4com_Acc when controlling ISI and RI*

		D5_WPicN_4com_Acc				
Model	Variable	*B*	*SEB*	*t*	*p*	η_p^2
Model 1	CVMT_ZAcqDelayed	0.038	0.019	2.049	0.045	0.062
Model 3	WM_ZShapeBOspan	0.056	0.018	3.111	0.003	0.133

Table 40. *Parameter Estimates of CVMT plus WM on D5_WPicN_4com_Acc when controlling ISI and RI*

		D5_WPicN_4com_Acc				
Model	Variable	*B*	*SEB*	*t*	*p*	η_p^2
Model 5	CVMT_ZAcqDelayed	0.022	0.019	1.174	0.245	0.022
	WM_ZShapeBOspan	0.048	0.019	2.559	0.013	0.096

6.6.2 Retention of tone production skill (as in oral word naming) on old words

As for the retention of tone production skill on practiced words as measured by oral word naming on old words, it was hypothesized that musical aptitude plays a facilitative role in tone production accuracy (Hypothesis 4b-i), and PM ability plays a facilitative role in oral word naming RT (i.e., the higher procedural memory ability, the faster word naming RT) when controlling for L1 word naming RT (Hypothesis 4b-ii).

Model 1 in Table 41 was conducted to test Hypothesis 4b-i. The results showed that musical aptitude was significantly related to tone production accuracy in oral word naming on old words, F (1, 63) = 5.475, p = .022, η_p^2 = .080, after controlling ISI and RI. Model 2 further tested whether musical aptitude interacted with treatment and found that musical aptitude neither interacted with ISI nor with RI (see Table 41), suggesting that the relationship between musical aptitude and the

retention of tone production accuracy did not differ depending on ISI or RI. The positive coefficient (B = 0.070) for musical aptitude on D5_OWN_old_Tone_Acc (see Table 42) suggests that musical aptitude plays a facilitative role on retention performance in tone production accuracy; therefore, Hypothesis 4b-i was confirmed.

Table 41. *Summary of ANCOVA models testing the effects of musical aptitude on D5_OWN_old_Tone_Acc when controlling ISI and RI*

Model 1	D5_OWN_old_Tone_Acc		
df (1, 63)	*F*	*p*	η_p^2
ISI	0.331	0.567	0.005
RI	0.582	0.448	0.009
ISI * RI	0.118	0.732	0.002
Music_Z 2PitchSTM 1PTM	5.475	0.022	0.080

Model 2	D5_OWN_old_Tone_Acc		
df (1, 61)	*F*	*p*	η_p^2
ISI	0.331	0.567	0.005
RI	0.539	0.466	0.009
ISI * RI	0.114	0.736	0.002
Music_Z 2PitchSTM 1PTM	4.572	0.037	0.070
ISI * Music_Z 2PitchSTM 1PTM	0.063	0.803	0.001
RI * Music_Z 2PitchSTM 1PTM	0.172	0.680	0.003

Table 42. *Parameter Estimates of Musical aptitude on D5_OWN_old_Tone_Acc when controlling ISI and RI*

		D5_OWN_old_Tone_Acc				
Model	Variable	*B*	*SEB*	*t*	*p*	η_p^2
Model 1	Music	0.070	0.030	2.340	0.022	0.080

To test Hypothesis 4b-ii, Model 3 (see Table 43) was conducted to see whether PM ability as measured by SRT plays a role on oral word naming RT when controlling for individual differences in L1 word naming RT and any effects of ISI

and RI. The results showed that SRT was not significantly related to oral word naming RT, F (1, 58) = 0.269, p = .606, η_p^2 = .005, while L1 word naming RT was a significant covariate, F (1, 58) = 12.848, p = .001, η_p^2 = .181. Model 4 was conducted to see whether the covariate L1 word naming RT interacted with experimental conditions. It turned out that the interaction between L1 word naming RT and ISI was nonsignificant, F (1, 56) = 1.147, p = .289, η_p^2 = .020, suggesting that the relationship between L1 word naming RT and oral Mandarin word naming RT on practiced words did not differ at different levels of ISI (i.e., daily vs. weekly practice schedule); however, the interaction between L1 word naming RT and RI was significant, F (1, 56) = 5.801, p = .019, η_p^2 = .094, suggesting that the relationship between L1 word naming RT and Mandarin word naming RT on old words differed at different levels of RI (i.e., 1 week vs 4 weeks). Therefore, the interaction between RI and L1 word naming RT was retained in further modeling, while the interaction between ISI and L1 word naming RT was not. Model 5 (see Table 43) was further conducted to see whether SRT interacted with ISI or RI. The results showed that the interaction between SRT and RI was nonsignificant, F (1, 55) = 1.207, p = .277, η_p^2 = .021, but the interaction between ISI and SRT was significant, F (1, 55) = 4.996, p = .029, η_p^2 = .083, suggesting that the relationship between SRT and oral word naming RT on practiced words did not differ depending on RI, but did differ depending on ISI. Another model, i.e., Model 6, was conducted to include the significant interactions between covariates and ISI or RI and exclude the nonsignificant interactions between covariates and experimental conditions (see Table 43). The parameter estimates on oral word naming RT on old words from Model 6 are presented in Table 44.

Table 43. *Summary of ANCOVA models testing the effects of PM ability on D5_OWN_old_RT_M_lg when controlling ISI, RI, and L1 word naming RT*

Model 3	D5_OWN_old_RT_M_lg		
df (1, 58)	*F*	*p*	η_p^2
ISI	1.091	0.300	0.018
RI	0.110	0.741	0.002
ISI * RI	0.890	0.349	0.015
EngOWN_RT_M_lg	12.848	0.001	0.181
SRT_Z	0.269	0.606	0.005

Model 4	D5_OWN_old_RT_M_lg		
df (1, 56)	*F*	*p*	η_p^2
ISI	1.095	0.300	0.019
RI	5.758	0.020	0.093
ISI * RI	1.438	0.236	0.025
EngOWN_RT_M_lg	12.942	0.001	0.188
SRT_Z	0.020	0.887	0.000
ISI * EngOWN_RT_M_lg	1.147	0.289	0.020
RI * EngOWN_RT_M_lg	5.801	0.019	0.094

Model 5	D5_OWN_old_RT_M_lg		
df (1, 55)	*F*	*p*	η_p^2
ISI	1.160	0.286	0.021
RI	5.385	0.024	0.089
ISI * RI	1.188	0.280	0.021
EngOWN_RT_M_lg	19.852	0.000	0.265
SRT_Z	0.039	0.845	0.001
RI * EngOWN_RT_M_lg	5.451	0.023	0.090
ISI * SRT_Z	4.996	0.029	0.083
RI * SRT_Z	1.207	0.277	0.021

Model 6	D5_OWN_old_RT_M_lg		
df (1, 56)	*F*	*p*	η_p^2
ISI	1.602	0.211	0.028
RI	6.010	0.017	0.097
ISI * RI	1.211	0.276	0.021
EngOWN_RT_M_lg	18.577	0.000	0.249
SRT_Z	0.003	0.956	0.000
RI * EngOWN_RT_M_lg	6.078	0.017	0.098
ISI * SRT_Z	4.626	0.036	0.076

Figure 16 presents the scatterplot and regression lines of D5_OWN_old_RT_M_lg against EngOWN_RT_M_lg at the two levels of RI. It shows that the relationship between L1 word naming RT and Mandarin word naming RT on practiced words was positive for both the short-term 1-week RI and the longer 4-week RI; however, the relationship seems to be much stronger at the shorter 1-week RI (R^2 = .469, thus R = .685) than at the longer 4-week RI (R^2 = .046, thus R = .214).

Table 44. *Parameter estimates on D5_OWN_old_RT_M_lg from Model 6*

Parameter	*B*	*SEB*	*t*	*p*	η_p^2
Intercept	1.876	.921	2.038	.046	.069
[ISI=0]	.065	.039	1.669	.101	.047
[ISI=1]	0[a]	.	.	.	.
[RI=0]	-3.322	1.365	-2.434	.018	.096
[RI=1]	0[a]	.	.	.	.
[ISI=0] * [RI=0]	-.061	.056	-1.101	.276	.021
[ISI=0] * [RI=1]	0[a]	.	.	.	.
[ISI=1] * [RI=0]	0[a]	.	.	.	.
[ISI=1] * [RI=1]	0[a]	.	.	.	.
EngOWN_RT_M_lg	.449	.337	1.329	.189	.031
SRT_Z	-.032	.021	-1.530	.132	.040
[RI=0] * EngOWN_RT_M_lg	1.232	.500	2.465	.017	.098
[RI=1] * EngOWN_RT_M_lg	0[a]	.	.	.	.
[ISI=0] * SRT_Z	.062	.029	2.151	.036	.076
[ISI=1] * SRT_Z	0[a]	.	.	.	.

a. This parameter is set to zero because it is redundant.

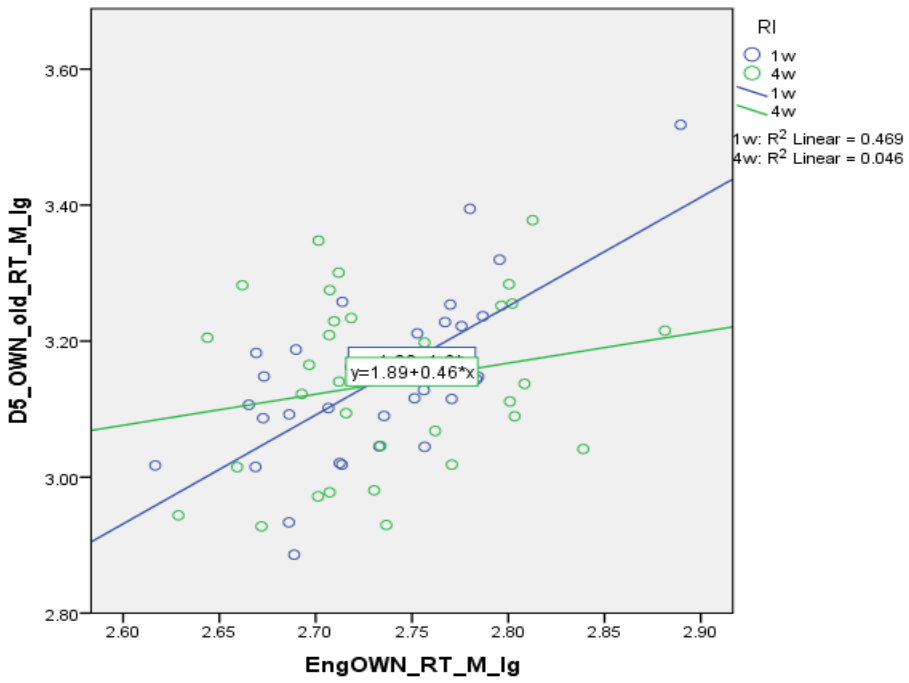


Figure 16. Scatterplot and regression lines of D5_OWN_old_RT_M_lg against EngOWN_RT_M_lg for each of the RI levels

Figure 17 presents the scatterplot and regression lines of D5_OWN_old_RT_M_lg against SRT_Z for each of the ISI levels. It shows that the relationship between PM ability as measured by SRT and oral word naming RT on practiced words was negative for the 1-week ISI (R^2=.024, thus R= - .155), suggesting that for the weekly practice groups, participants with higher PM ability tended to respond faster in oral word naming on practiced words on Day 5, irrespective of RI. On the contrary, the relationship between PM ability and oral word naming RT on practiced words was positive for the 1-day ISI (R^2=.039, R= .197), suggesting that for the daily practice groups, participants with higher PM ability tended to respond slower in oral word naming on practiced words on Day 5, irrespective of RI. This interaction between SRT and ISI suggests that the relationship between PM ability and the retention of oral word naming automaticity is more complex than a simple facilitative role as hypothesized. These results will be discussed in the discussion section.

Figure 18 presents the scatterplot and regression lines of D5_OWN_old_RT_M_lg against SRT_Z for each of the four experimental groups. Descriptively, it seems that there were no relationship between SRT and oral word naming RT on practiced words in Groups A and C, the two short 1-week RI groups. In other words, the short 1-week RI does not seem to draw on PM ability for the retention of oral tone production automaticity. The relationship between SRT and oral word naming RT on practiced words seems to be drastically different, in fact, opposite for Group D (weekly practice with long RI) and Group B (daily practice with long RI). For the long-term 4-week RI, stronger PM ability seems to speed up

oral word naming RT (on old words) in the distributed practice group (Group D), but slow down oral word naming RT (on old words) in the massed practice group (Group B).

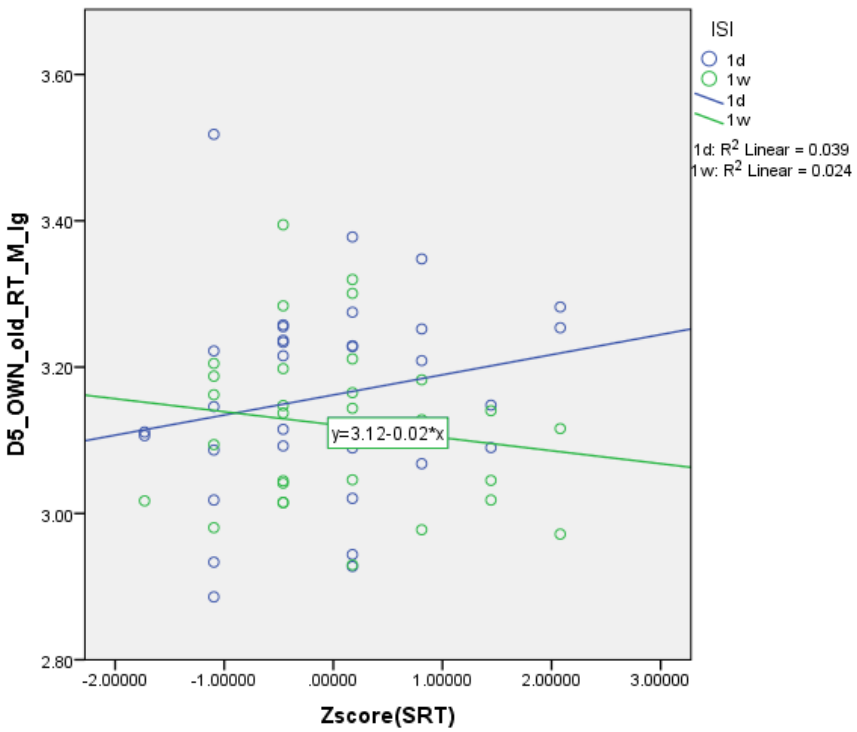


Figure 17. Scatterplot and regression lines of D5_OWN_old_RT_M_lg against SRT_Z for each of the ISI levels

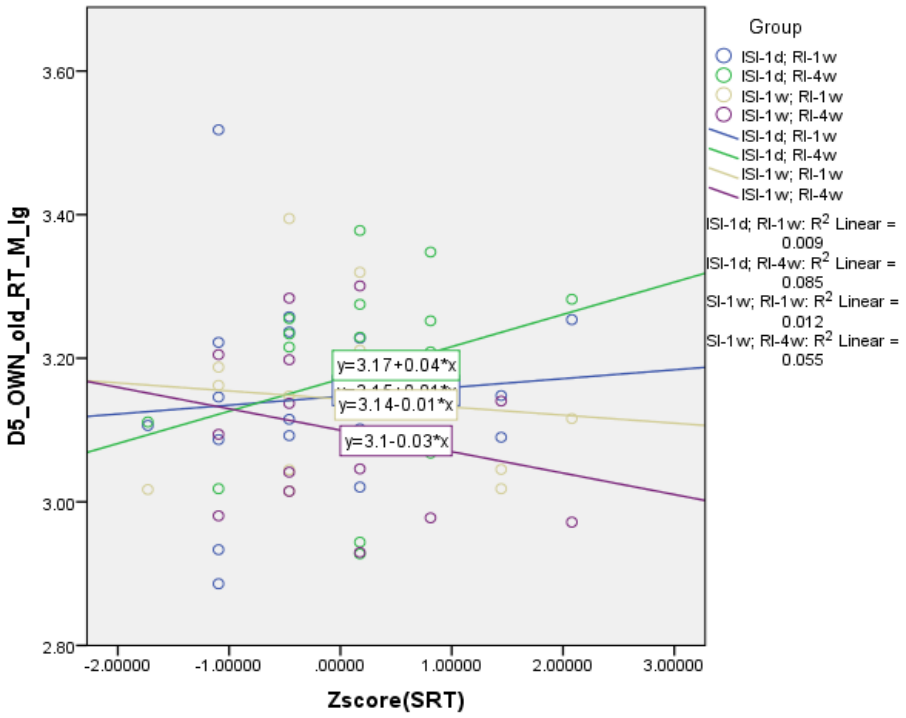


Figure 18. Scatterplot and regression lines of D5_OWN_old_RT_M_lg against SRT_Z for each of the four experimental groups

6.6.3 Retention of tone production skill (as in oral word naming) for new words

With regard to the retention of tone production skill in new words, as measured by the oral word naming task on new words, it was hypothesized that WM capacity and musical aptitude each play a facilitative role in tone production accuracy on new words (Hypothesis 4c-i, Hypothesis 4c-ii), and PM ability plays a facilitative role in oral word naming RT on new words (i.e., the higher procedural memory ability, the faster word naming RT) when controlling for L1 word naming RT (Hypothesis 4c-iii).

Model 1 in Table 45 was conducted to test Hypothesis 4c-i regarding the role of WM capacity on tone production accuracy on new words. The results showed that after controlling for ISI and RI, WM capacity turned out to be significantly related to

tone production accuracy in oral word naming of new words, F (1, 63) = 9.398, p = .003, with large effect size, η_p^2 = .130. Model 2 further tested whether WM interacted with treatment and found that WM neither interacted with ISI nor with RI (see Table 45), suggesting that the relationship between WM capacity and retention of tone production accuracy on new words did not differ depending on either ISI or RI. The positive coefficient (B = 0.077) for WM on D5_OWN_new_Tone_Acc (see Table 46) suggests that WM capacity plays a facilitative role on the retention of tone production accuracy in new words, confirming Hypothesis 4c-i.

Table 45. *Summary of ANCOVA models testing the effects of aptitudes on D5_OWN_new_Tone_Acc when controlling ISI and RI*

Model 1	D5_OWN_new_Tone_Acc			Model 3	D5_OWN_new_Tone_Acc		
df (1, 63)	F	p	η_p^2	df (1, 63)	F	p	η_p^2
ISI	1.580	0.213	0.024	ISI	0.716	0.401	0.011
RI	2.818	0.098	0.043	RI	3.193	0.079	0.048
ISI * RI	0.210	0.648	0.003	ISI * RI	0.129	0.721	0.002
WM_Z ShapeBOspan	9.398	0.003	0.130	Music_Z 2PitchSTM1PTM	6.443	0.014	0.093

Model 2	D5_OWN_new_Tone_Acc			Model 4	D5_OWN_new_Tone_Acc		
df (1, 61)	F	p	η_p^2	df (1, 61)	F	p	η_p^2
ISI	1.615	0.209	0.026	ISI	0.667	0.417	0.011
RI	2.785	0.100	0.044	RI	3.106	0.083	0.048
ISI * RI	0.240	0.626	0.004	ISI * RI	0.125	0.724	0.002
WM_Z ShapeBOspan	9.583	0.003	0.136	Music_Z 2PitchSTM1PTM	6.309	0.015	0.094
ISI * WM_Z ShapeBOspan	0.301	0.585	0.005	ISI * Music_Z 2PitchSTM1PTM	0.144	0.706	0.002
RI * WM_Z ShapeBOspan	0.214	0.645	0.004	RI * Music_Z 2PitchSTM1PTM	0.013	0.911	0.000

Model 3 in Table 45 was run to test Hypothesis 4c-ii regarding the role of musical aptitude on tone production accuracy in oral word naming of new words. The

results from Model 3 showed that after controlling for ISI and RI, musical aptitude turned out to be significantly related to oral production accuracy in new words, F (1, 63) = 6.443, p = .014, η_p^2 = .093. Model 4 further tested whether musical aptitude interacted with experimental conditions, and found that musical aptitude did not interact with either ISI or RI (see Table 45), suggesting that the relationship between musical aptitude and tone production accuracy in oral word naming of new words did not differ depending on ISI or RI. The positive coefficient (B = 0.073) for musical aptitude on tone production accuracy on new words (see Table 46) suggests that participants with higher musical aptitude performed better in tone production accuracy on new words, therefore confirming Hypothesis 4c-ii.

Table 46. *Parameter Estimates of Aptitudes in Single-Covariate Analyses on D5_OWN_new_Tone_Acc when controlling ISI and RI*

		D5_OWN_new_Tone_Acc				
Model	Variable	*B*	*SEB*	*t*	*p*	η_p^2
Model 1	WM_ZShapeBOspan	0.077	0.025	3.066	0.003	0.130
Model 3	Music_Z2PitchSTM1PTM	0.073	0.029	2.538	0.014	0.093

Model 5 in Table 47 was conducted to test Hypothesis 4c-iii regarding the role of PM ability on oral word naming RT of new words. The results showed that, when controlling for L1 word naming RT and any effects of ISI and RI, SRT was not significantly related to oral word naming RT on new words, F (1, 61) = 1.101, p = .298, η_p^2 = .018. L1 word naming RT was not a significant covariate, either, F (1, 61) = 2.083, p = .154, η_p^2 = .033. Model 6 was run to check whether L1 word naming RT interacted with ISI or RI; neither of the interactions turned out to be significant (see Table 47). Model 7 was run to check whether SRT interacted with ISI or RI, and

neither interaction turned out to be significant, either (see Table 47). Table 48 presents the parameter estimates of L1 word naming RT and SRT on oral word naming RT on new words. The results disconfirmed Hypothesis 4c-iii; PM ability as measured by SRT did not seem to play a role in oral word naming RT on new words.

Table 47. *Summary of ANCOVA models testing the effects of PM ability on D5_OWN_new_RT_M_lg when controlling for ISI, RI, and L1 word naming RT*

Model 5	D5_OWN_new_RT_M_lg		
df (1, 61)	*F*	*p*	η_p^2
ISI	0.234	0.630	0.004
RI	0.034	0.855	0.001
ISI * RI	0.234	0.630	0.004
EngOWN_RT_M_lg	2.083	0.154	0.033
SRT_Z	1.101	0.298	0.018

Model 6	D5_OWN_new_RT_M_lg		
df (1, 59)	*F*	*p*	η_p^2
ISI	0.151	0.699	0.003
RI	2.080	0.155	0.034
ISI * RI	0.405	0.527	0.007
EngOWN_RT_M_lg	2.064	0.156	0.034
SRT_Z	0.571	0.453	0.010
ISI * EngOWN_RT_M_lg	0.160	0.691	0.003
RI * EngOWN_RT_M_lg	2.093	0.153	0.034

Model 7	D5_OWN_new_RT_M_lg		
df (1, 59)	*F*	*p*	η_p^2
ISI	0.255	0.615	0.004
RI	0.160	0.691	0.003
ISI * RI	0.088	0.768	0.001
EngOWN_RT_M_lg	3.112	0.083	0.050
SRT_Z	1.209	0.276	0.02
ISI * SRT_Z	3.499	0.066	0.056
RI * SRT_Z	0.118	0.732	0.002

Table 48. *Parameter Estimates of PM ability on D5_OWN_new_RT_M_lg when controlling ISI, RI, and L1 word naming RT*

		D5_OWN_new_RT_M_lg				
Model	Variable	*B*	*SEB*	*t*	*p*	η_p^2
Model 5	EngOWN_RT_M_lg	0.601	0.417	1.443	0.154	0.033
	SRT_Z	0.025	0.024	1.049	0.298	0.018

6.6.4 Retention of oral word production skill (as in oral picture naming)

With regard to the retention of oral Mandarin word production skill as measured by the oral picture naming task, it was hypothesized that DM ability, WM capacity, PSTM, and musical aptitude each independently play a facilitative role in oral picture naming accuracy (Hypotheses 4d-i, 4d-ii, 4d-iii, 4d-iv), and PM ability plays a facilitative role in oral picture naming RT (i.e., the higher procedural memory ability, the faster word naming RT) when controlling for L1 word naming RT (Hypothesis 4d-v).

Model 1 presented in Table 49 tests Hypothesis 4d-i regarding the role of DM ability on retention performance in oral picture naming accuracy. The results showed that when controlling the effects of ISI and RI, DM ability as measured by CVMT was not significantly related to oral picture naming accuracy on the retention test, F (1, 63) = 2.053, p = .157, η_p^2 = .032. Model 2 further checked whether CVMT interacted with ISI or RI, and found that neither interaction was statistically significant (see Table 49). The parameter estimates for CVMT are presented in Table 50. These results suggest that DM ability did not play a significant role on the retention performance in oral picture naming accuracy, therefore disconfirming Hypothesis 4d-i.

Model 3 in Table 49 tests Hypothesis 4d-ii regarding the role of WM capacity on retention performance in oral picture naming accuracy. The results from Model 3 showed that after controlling the effects of ISI and RI, WM capacity was still significantly related to oral picture naming accuracy on Day 5, F (1, 63) = 4.445, p = .039, η_p^2 = .066. Model 4 further tested whether WM interacted with the

experimental conditions, and found that WM did not interact with either ISI or RI (see Table 49), suggesting that the relationship between WM capacity and retention performance in oral picture naming accuracy did not differ depending on ISI or RI. The positive coefficient (B = 0.039) of WM on D5_OPicN_4com_Acc (see Table 50) suggests a facilitative role of WM capacity on retention performance in oral picture naming accuracy, thus confirming Hypothesis 4d-ii.

Table 49. *Summary of ANCOVA models testing the effects of aptitudes on D5_OPicN_4com_Acc when controlling ISI and RI*

Model 1	D5_OPicN_4com_Acc		
df (1, 63)	F	p	η_p^2
ISI	0.569	0.454	0.009
RI	16.806	0.000	0.211
ISI * RI	2.649	0.109	0.040
CVMT_Z AcqDelayed	2.053	0.157	0.032

Model 3	D5_OPicN_4com_Acc		
df (1, 63)	F	p	η_p^2
ISI	0.269	0.606	0.004
RI	15.631	0.000	0.199
ISI * RI	2.622	0.110	0.040
WM_Z ShapeBOspan	4.445	0.039	0.066

Model 2	D5_OPicN_4com_Acc		
df (1, 61)	F	p	η_p^2
ISI	0.594	0.444	0.010
RI	17.149	0.000	0.219
ISI * RI	2.442	0.123	0.038
CVMT_Z AcqDelayed	2.870	0.095	0.045
ISI * CVMT_Z AcqDelayed	0.488	0.487	0.008
RI * CVMT_Z AcqDelayed	0.955	0.332	0.015

Model 4	D5_OPicN_4com_Acc		
df (1, 61)	F	p	η_p^2
ISI	0.214	0.645	0.003
RI	15.572	0.000	0.203
ISI * RI	2.049	0.157	0.032
WM_Z ShapeBOspan	4.871	0.031	0.074
ISI * WM_Z ShapeBOspan	0.031	0.861	0.001
RI * WM_Z ShapeBOspan	1.647	0.204	0.026

Model 5	D5_OPicN_4com_Acc		
df (1, 63)	F	p	η_p^2
ISI	0.567	0.454	0.009
RI	15.283	0.000	0.195
ISI * RI	2.537	0.116	0.039
NWR_Z	0.004	0.947	0.000

Model 7	D5_OPicN_4com_Acc		
df (1, 63)	F	p	η_p^2
ISI	0.642	0.426	0.010
RI	16.491	0.000	0.207
ISI * RI	2.843	0.097	0.043
Music_Z 2PitchSTM1PTM	4.548	0.037	0.067

Model 6	D5_OPicN_4com_Acc		
df (1, 61)	*F*	*p*	η_p^2
ISI	0.728	0.397	0.012
RI	13.975	0.000	0.186
ISI * RI	2.692	0.106	0.042
NWR_Z	0.006	0.938	0.000
ISI * NWR_Z	0.069	0.793	0.001
RI * NWR_Z	0.298	0.587	0.005

Model 8	D5_OPicN_4com_Acc		
df (1, 61)	*F*	*p*	η_p^2
ISI	0.665	0.418	0.011
RI	16.329	0.000	0.211
ISI * RI	2.782	0.100	0.044
Music_Z 2PitchSTM1PTM	4.875	0.031	0.074
ISI * Music_Z 2PitchSTM1PTM	0.337	0.564	0.005
RI * Music_Z 2PitchSTM1PTM	0.413	0.523	0.007

Model 5 in Tale 49 tests Hypothesis 4d-iii concerning the role of PSTM on retention performance in oral picture naming accuracy when controlling for ISI and RI. The results showed that PSTM was not significantly related to oral picture naming accuracy on the retention test, F (1, 63) = 0.004, p = .947, $\eta_p^2 < .001$. Model 6 further tested whether PSTM interacted with the experimental conditions, and found that PSTM did not interact with either ISI or RI (see Table 50). These results suggest that PSTM did not play a role on the retention performance in oral picture naming accuracy, therefore disconfirming Hypothesis 4d-i.

Model 7 in Table 49 tests Hypothesis 4d-iv regarding the role of musical aptitude on the retention performance in oral picture naming accuracy. The results showed that after controlling for ISI and RI, musical aptitude was still significantly related to oral picture naming accuracy on the retention test, F (1, 63) = 4.548, p = .037, η_p^2 = .067. Model 8 further tested whether musical aptitude interacted with the experimental conditions, and found that musical aptitude did not interact with either ISI or RI, suggesting that the relationship between musical aptitude and retention performance in oral picture naming did not differ depending on ISI or RI. The

positive coefficient (B = 0.044) of musical aptitude on oral picture naming accuracy on Day 5 suggests a facilitative role of musical aptitude on retention performance in oral picture naming accuracy, and thus confirming Hypothesis 4d-iv.

Table 50. *Parameter Estimates of Aptitudes in Single-Covariate Analyses on D5_OPicN_4com_Acc when controlling ISI and RI*

		D5_OPicN_4com_Acc				
Model	Variable	*B*	*SEB*	*T*	*P*	η_p^2
Model 1	CVMT_ZAcqDelayed	0.027	0.019	1.433	0.157	0.032
Model 3	WM_ZShapeBOspan	0.039	0.018	2.108	0.039	0.066
Model 5	NWR_Z	-0.001	0.017	-0.067	0.947	0.000
Model 7	Music_Z2PitchSTM1PTM	0.044	0.020	2.133	0.037	0.067

With regard to the role of PM ability as measured by SRT on the retention performance in oral picture naming RT (Hypothesis 4d-v), three ANCOVA models were conducted (see Table 51). The results from Model 9 showed that after controlling the effects of ISI and RI and individual differences in L1 word naming RT, SRT was not significantly related to oral picture naming RT on Day 5, F (1, 55) = 1.458, p = .232, η_p^2 = .026. On the other hand, L1 word naming RT turned out to be a significant covariate, F (1, 55) = 10.074, p = .002, with large effect size, η_p^2 = .155. Model 10 was conducted to further check whether L1 word naming RT interacted with the experimental conditions, and found that L1 word naming RT did not interact with either ISI or RI on oral picture naming RT (see Table 51). Model 11 was conducted to see whether SRT interacted with the experimental conditions. The results yielded a nonsignificant interaction between ISI and SRT, F (1, 53) = 3.057, p = .086, η_p^2 = .055, but the interaction between RI and SRT turned out to be statistically significant, F (1, 53) = 4.531, p = .038, η_p^2 = .079. These results suggest

that the relationship between SRT and the retention performance in oral picture naming RT differed significantly on different levels of RI, but it did not differ substantially depending on ISI. The parameter estimates on oral picture naming RT from Model 11 are presented in Table 52.

Table 51. *Summary of ANCOVA models testing the effects of PM ability on D5_OPicN_RT_M_lg when controlling ISI, RI, and L1 word naming RT*

Model 9	D5_OPicN_RT_M_lg		
df (1, 55)	*F*	*p*	η_p^2
ISI	10.157	0.002	0.156
RI	4.322	0.042	0.073
ISI * RI	0.450	0.505	0.008
EngOWN_RT_M_lg	10.074	0.002	0.155
SRT_Z	1.458	0.232	0.026

Model 10	D5_OPicN_RT_M_lg		
df (1, 53)	*F*	*p*	η_p^2
ISI	0.023	0.880	0.000
RI	0.244	0.623	0.005
ISI * RI	0.478	0.492	0.009
EngOWN_RT_M_lg	8.580	0.005	0.139
SRT_Z	1.225	0.273	0.023
ISI * EngOWN_RT_M_lg	0.045	0.832	0.001
RI * EngOWN_RT_M_lg	0.207	0.651	0.004

Model 11	D5_OPicN_RT_M_lg		
df (1, 53)	*F*	*P*	η_p^2
ISI	9.663	0.003	0.154
RI	4.123	0.047	0.072
ISI * RI	0.444	0.508	0.008
EngOWN_RT_M_lg	14.355	0.000	0.213
SRT_Z	1.668	0.202	0.031
ISI * SRT_Z	3.057	0.086	0.055
RI * SRT_Z	4.531	0.038	0.079

Table 52. *Parameter estimates on D5_OPicN_RT_M_lg from Model 11*

Parameter	*B*	*SEB*	*t*	*p*	η_p^2
Intercept	-.572	1.058	-.541	.591	.005
[ISI=0]	.158	.058	2.719	.009	.122
[ISI=1]	0[a]	.	.	.	.
[RI=0]	-.057	.062	-.912	.366	.015
[RI=1]	0[a]	.	.	.	.
[ISI=0] * [RI=0]	-.056	.084	-.666	.508	.008
[ISI=0] * [RI=1]	0[a]	.	.	.	.
[ISI=1] * [RI=0]	0[a]	.	.	.	.
[ISI=1] * [RI=1]	0[a]	.	.	.	.
EngOWN_RT_M_lg	1.468	.387	3.789	.000	.213
SRT_Z	.039	.042	.937	.353	.016
[ISI=0] * SRT_Z	.081	.046	1.748	.086	.055
[ISI=1] * SRT_Z	0[a]	.	.	.	.
[RI=0] * SRT_Z	-.099	.047	-2.129	.038	.079
[RI=1] * SRT_Z	0[a]	.	.	.	.

a. This parameter is set to zero because it is redundant.

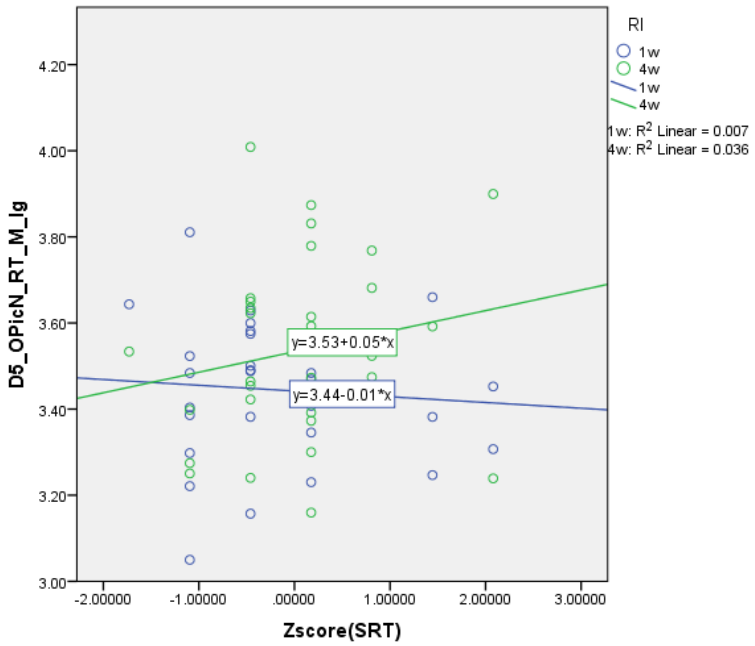


Figure 19. Scatterplot and regression lines of D5_OPicN_RT_M_lg against SRT_Z for each of the RI levels

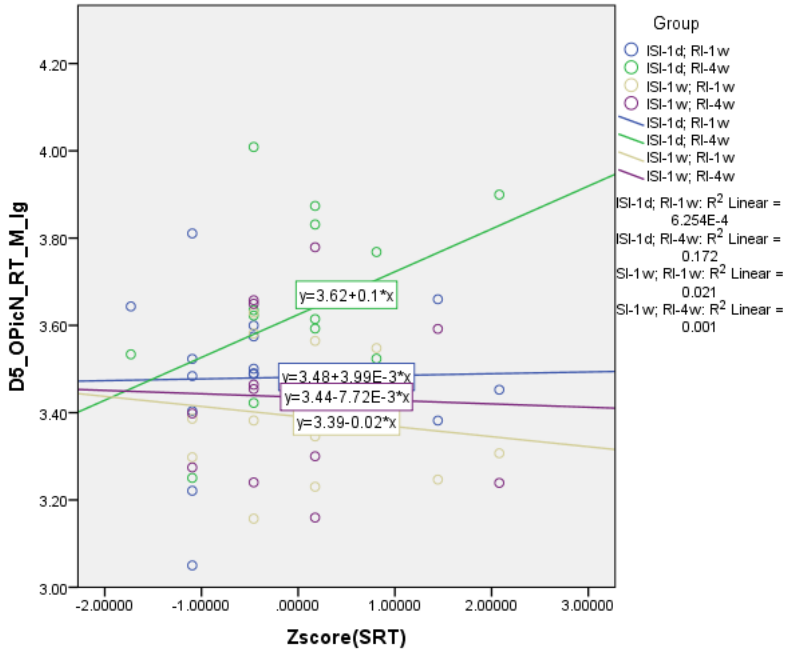


Figure 20. Scatterplot and regression lines of D5_OPicN_RT_M_lg against SRT_Z for each of the four experimental conditions

Figure 19 presents the scatterplot and regression lines of D5_OPicN_RT_M_lg against SRT_Z for each of the RI levels. It shows that the relationship between PM ability as measured by SRT and oral picture naming RT was almost negligible for the 1-week RI (R^2 = .007, thus R = - .084). For the longer 4-week RI, the relationship between SRT and oral picture naming RT on Day 5 seems to be positive, though not statistically significant (R^2 =.036, thus R = .190), suggesting that participants with higher PM ability tended to respond slower in oral picture naming on the retention test. These results disconfirmed Hypothesis 4d-v, which anticipated a negative correlation between SRT scores and oral picture naming RTs.

Figure 20 presents the scatterplot and regression lines of D5_OPicN_RT_M_lg against SRT_Z for each of the four experimental conditions. It shows that the regression lines for Group A and Group D are pretty flat, which indicates no relationship between PM ability and oral picture naming RT in these two groups. The regression line for Group C (distributed practice with a short 1-week RI) is the only one showing a negative correlation, though not statistically significant (R^2 = .021, thus R = - .145) (see Appendix N also). On the other hand, for Group B (massed practice with a longer 4-week RI), the regression line seems to indicate a positive correlation (R^2 = .172, thus R = .415); however the correlation was not statistically significant either (see Appendix N). These results will be discussed in the discussion section.

Chapter 7: Discussion

The present dissertation investigated the effects of temporal distribution of practice (ISI: 1-day vs. 1-week) on the automatization and retention (RI: 1-week vs. 4-week) of L2 Mandarin word production by a group of naïve native speakers of English. This study examined the effects of ISI on the acquisition of oral Mandarin tonal word production skill across the training sessions. More importantly, this study attempted to explore whether the effects of ISI and RI differ depending on (a) the type of knowledge/skill to be retained, i.e., purely declarative word knowledge vs. skills in oral production, and (b) the proportion of declarative knowledge required to perform the skill for skill retention. Finally, this study scrutinized the roles of cognitive aptitudes (including working memory, phonological STM, declarative memory, procedural memory, and musical aptitude) at different stages of learning oral Mandarin word production and on the retention of the learned knowledge or skills.

Eighty native English-speaking adults who did not have any prior knowledge of a tonal language completed all sessions of the study and provided data for analyses. These participants were randomly assigned to four experimental conditions, i.e., Condition A with an ISI of 1 day and an RI of 1 week (ISI/RI: 14%), Condition B with an ISI of 1 day and an RI of 4 weeks (4%), Condition C with an ISI of 1 week and an RI of 1 week (100%), and Condition D with an ISI of 1 week and an RI of 4 weeks (25%). Each participant came in for five sessions. The four groups completed a set of cognitive aptitude tests and underwent the same number and content of sessions, which differed only on training or testing schedules.

All participants completed a pretest in oral word naming of the to-be-learned target words at the beginning of the first disyllabic word learning session. An oral picture-naming task was administered as a post-session quiz at the end of the first training session, and then as a pre- or post-session quiz at the beginning and the end of the second and third training sessions to keep track of learning within training sessions or forgetting/degradation between the gaps. Three outcome tasks were administered as the retention test (or delayed posttest) in the last session after a retention interval: an oral picture naming task that taps the global oral word production skill from meaning to articulation; a written picture naming task that measures the declarative component of word knowledge (i.e., picture-Pinyin mappings); and an oral word naming task that measures oral tone production skill from Pinyin to articulation in both old/practiced words and new words.

7.1 Temporal Distribution of Practice on the automatization and retention of L2 Mandarin word production

This study examined both the effects of ISI during the course of acquisition of oral Mandarin word production skill, and the effects of ISI and RI on the retention of the learned knowledge or skill. The results will be summarized and discussed in two subsections.

7.1.1 Effects of ISI on skill acquisition

Table 53 summarizes the results regarding the effects of ISI on oral picture naming performance (the global oral word production skill from meaning to articulation) at the beginning and the end of the second and third training sessions. It

was found that ISI only had an effect on the *pre*-session performance in oral picture naming accuracy at the beginning of TS2 and TS3, but did not show effects on the post-session performance at the end of the two training sessions, in terms of either accuracy (at the end of both TS2 and TS3) or RT (at the end of TS3; RT measures at the earlier stages, i.e., before the end of TS3, were not generated due to high error rates). These results, therefore, refuted both Hypothesis 1a and 1b for an advantage for massed practice over distributed across the board in all five outcome measures.

Table 53. *Summary of the effects of ISI on performance across the training sessions*

	TS2		TS3		
	Pre_OPicN	Post_OPicN	Pre_OPicN	Post_OPicN	
	Acc	Acc	Acc	Acc	RT
ISI	√	x	√	x	x

Note. √ means a statistically significant effect was found; x means a statistically significant effects was not found; significance was set at $p < .05$.

It was found that the short 1-day ISI resulted in much less forgetting or degradation in oral picture naming accuracy between sessions than the longer 1-week ISI. This is logical because longer gaps tend to lead to more forgetting, especially since knowledge by the end of the first training session largely remained at the declarative level, and declarative knowledge is vulnerable to forgetting. The short 1-day ISI group, or the massed practice group, was thus advantaged at the beginning of TS2 and TS3; this advantage, however, did not hold through to the end of the training sessions. This was not as hypothesized; however, this result was not surprising either. The advantage of the massed practice group at the beginning of the training sessions was apparently overshadowed by the effectiveness of training. In TS2 and TS3, before participants started to engage in oral production practice, they were asked to

review declarative knowledge about the tones and were given feedback. This review step may have helped the longer ISI group, or the distributed practice group to a greater extent, because this knowledge may still be fresh for the daily training group but might have been forgotten to some degree to the weekly training group. Immediately after the review step, participants went to another intensive oral practice session with feedback given to each trial of practice. Thus, although the distributed group started lower due to more forgetting or skill decay at the beginning of the training sessions, the subsequent training was effective enough (in terms of both quality and quantity) to pull them to the same level as the massed practice group was able to reach.

This finding seems to be inconsistent with the robust advantage Suzuki (2017) found for massed practice (ISI-3.3day) over distributed practice (ISI-7day) in automatization of L2 morphology. It was found in Suzuki (2017) that the 3.3-day ISI group started to provide more accurate performance than the 7-day ISI group from the beginning of the third training session (i.e., pre- and post-session quizzes at the beginning and the end of TS3 and TS4), and this advantage was maintained on both the 1-week and 4-week delayed posttests. The findings of the two studies are parallel in that both found an advantage for the massed group in the pre-session quizzes; the findings of the two studies differ, however, in that the advantage was still present at the post-session quizzes in Suzuki (2017) but not in the present study. This inconsistent finding may be explained by a number of procedure differences. First of all, the training in Suzuki (2017) was completely computer-delivered, while the training in the present study was half computer-delivered, half-human delivered. Both

studies had a review step (review of information about the target structure) before participants started to engage in practice in the subsequent training sessions; however, the two studies differ in how the review step was delivered. In Suzuki (2017), participants were given one minute to review the explicit instruction sheet in Training Sessions 2-4 before starting to engage in grammar practice. In the present study, participants were asked to verbally describe the four Mandarin tones and the tonal change rule in disyllabic words to the experimenter, and received feedback from the experimenter. The active recall/retrieval processes required by the review step in the present study plus the interactional context may be more effective than the review procedure in Suzuki (2017), in which participants only passively or receptively reviewed the instruction. Secondly, the two studies differ in the procedure of delivering feedback. In Suzuki (2017), feedback was computer-delivered and therefore the feedback was not customized to the participant's responses. It seems from the report that all participants received the same feedback in the form of a correct response after each practice item. In the current study, customized feedback to the oral production for each trial during practice was delivered by the human experimenter. The feedback included a verbal judgment from the experimenter ("yes, correct", or "no", "close"; if incorrect, an identification of the problem) and an audio playing of the model pronunciation of the practiced word. Participants were also asked to imitate the model pronunciation immediately after hearing it. The customized feedback delivered through human interaction may be more effective than the same feedback delivered through computer. These two procedure differences may have rendered the training procedure in the present study more effective, which then

contributed to leveling up the performance of the initially disadvantaged distributed group (due to more forgetting) to the same level at the end of the training session as the massed practice group was able to reach. If the present study used the same procedure in conducting the review and delivering feedback, I would expect the study to yield the same results as what Suzuki (2017) found.

7.1.2 Effects of ISI and RI on skill retention (including subcomponents)

As has been discussed earlier, the target skill of oral production of Mandarin words from meaning to articulation as measured in oral picture naming consists of two subcomponents, i.e., meaning-word mappings and oral articulation of the words, which were measured by the written picture naming task and the oral word naming task, respectively. While the first subcomponent of meaning-word mappings is declarative in nature, the second subcomponent of oral articulation of the words is more of a skill. In the context of this study, the focus of the word production skill was oral tone production as the segments of the words were deliberately chosen to be easy for English NSs to pronounce. By design, a comparison of the results from the three outcome tasks, one tapping the global complex skill and two tapping the two subcomponents respectively, can enable us to see how the effects of ISI and RI differ or parallel on the forgetting/retention of the complex skill (declarative knowledge + skill) and of its subcomponents that differ in nature (declarative knowledge vs skill).

Table 54 presents the summary of the effects of ISI and RI on retention performance in the three tasks on the seven outcome measures, i.e., written picture naming accuracy, oral word naming accuracy and RT on old/practiced words, oral

word naming accuracy and RT on new words, and oral picture naming accuracy and RT. I will first discuss the results on the outcome measures and then compare them.

Table 54. *Summary of the effects of ISI and RI on retention performance*

	WPicN	OWN_old		OWN_new		OPicN	
	Acc	Acc	RT	Acc	RT	Acc	RT
ISI	x	x	x	√	x	x	√
RI	√	x	x	x	x	√	√
ISI*RI	x	x	x	x	x	x	x

Note. √ means a statistically significant effect was found; x means a statistically significant effects was not found; significance was set at $p < .05$.

Written picture naming accuracy was used to measure the retention of declarative knowledge in meaning-spelling mappings (including tone marking). The hypothesized interaction between ISI and RI was not found. It was not the case that 1-day ISI combined with 1-week RI group (Group A) and 1-week ISI combined with 4-week RI group (Group D), which have the optimal ISI/RI ratios, outperformed the other two groups with suboptimal ISI/RI ratios. Therefore, Hypothesis 3a was refuted. There was no significant main effect for ISI, either. Instead, a robust main effect for RI was found - Groups A and C with short 1-week RI were found to significantly outperform Groups B and D, the groups with the long 4-week RI. These results seem to suggest that retention of declarative knowledge is only determined by RI, and not by ISI: the longer the RI, the worse the retention of declarative knowledge. Although the interaction between ISI and RI did not turn out to be statistically significant, probably due to the relatively small sample size in each group, follow-up pairwise comparisons were conducted to see the simple main effect of ISI at each level of RI. It was found that for the short-term 1-week RI, the effect of ISI is negligible, F (1, 66)

= 0.203, p = .654, η_p^2=.003; for the long-term 4-week RI, however, ISI seems to play a role, F (1, 66) = 5.259, p = .025, with medium effect size, η_p^2=.074. The 1-week ISI as compared to the 1-day ISI improved retention of declarative knowledge at the longer 4-week RI. Taken together, the results of the present study do not provide support for Cepeda et al.'s (2006) and Rohrer and Pashler's (2007) theory about the optimal range of ISI/RI ratios (10% to 30%) for best retention. Instead, a robust RI effect was found, i.e., the longer the RI, the worse the retention performance in recalling declarative knowledge. While spacing did not exert an effect for the short RI, the effect of spacing was observed for the long RI; this pattern of results is in line with Cepeda et al.'s (2006) prediction that the effect of ISI depends on RI.

For skill retention, oral word naming accuracy and RT on old words were used to measure the retention of the highly automatized skill that participants went through three intensive training sessions to practice. The results suggest that neither ISI, nor RI (including the interaction between them) exerted an effect on the retention performance in either oral tone production accuracy or oral word naming RT on practiced words. The four groups performed at the same level in terms of both oral tone production accuracy and oral word naming RT on practiced words in the retention test. The longer 4-week RI did not result in more decay in either accuracy or RT than 1-week RI, and massed practice (1-day ISI) did not result in better retention performance than distributed practice (1-week ISI) either. Therefore, the hypotheses purporting an advantage of massed practice over distributed practice on the retention of this highly automatized skill in oral tone production of old words were not borne out in the current data. Looking at this skill decay longitudinally, recall that the four

groups were trained to the same high level in tone production accuracy in oral picture naming by the end of the last training session (TS3) (see details in Section 6.2, Table 17), and that the tone production accuracy rate drop from the end of TS3 to the final retention test Day 5 was almost negligible for the 1-week RI groups (i.e., .02 for Group A and less than .01 Group C), and very mall for the 4-week RI groups (i.e., .05 for Group B, and .09 for Group D). In other words, there seems to be little skill decay in tone production accuracy from at the end of the last training session to Day 5 after an RI.

These findings regarding the effects of ISI and RI on retention performance in written picture naming (declarative) and oral word naming on old words (highly automatized) parallel the findings of Paik and Ritter (2016), who examined how learning schedules (distributed, massed) interact with knowledge types (declarative, procedural) on retention. Using a Japanese-English vocabulary learning task as the declarative task, and a Tower of Hanoi task as the procedural task, Paik and Ritter (2016) found that there were noticeable losses in terms of both accuracy and RT in the declarative knowledge task from the last practice session to the retention session (RI = 3 weeks), performance in the last practice session was almost the same as in the retention session for the procedural task; in other words, the procedural knowledge did not decay over a retention interval of three weeks. The current study found the same pattern of results for the retention of declarative word knowledge (meaning-word mappings) versus oral tone production skill retention, in that while declarative word knowledge is vulnerable to memory decay (longer RI, worse retention), highly automatized skill in oral tone production was robust to delay (very little accuracy rate

drop in tone production). This pattern of results found in the present study and that of Paik and Ritter's (2016) suggests that procedural knowledge or highly automatized skill is much more robust and much less susceptible to decay than declarative knowledge, providing empirical support to one of the most important premises of skill retention theory (Kim et al., 2013).

Moving on to oral word naming of new words, no strong hypothesis was put forth regarding the effects of ISI and RI on oral tone production performance in new linguistic contexts, because no previous research has looked at these effects on the generalization of practiced skills in new contexts. As orally naming new Mandarin words from Pinyin involves fusing tone production with new combinations of segments, declarative knowledge about how to pronounce the tones was likely to be involved in this fusing process. In other words, this task involves the retention of both declarative knowledge and procedural knowledge about how to pronounce tones. Due to the double constraints for retention of declarative knowledge and retention of procedural knowledge, Group A was expected to outperform the other three groups in oral tone production accuracy on new words at least descriptively. This was indeed found to be true descriptively according to group means.

More importantly, a significant main effect for ISI was found on tone production accuracy in oral word naming of new words. That is, massed practice led to significantly better retention performance in oral tone production accuracy in new words than distributed practice (see Figure 12). In other words, the advantage of massed practice showed up on tone production accuracy on NEW words (such effect was hypothesized on tone production accuracy on old/practiced words but was not

borne out there). The reason why the advantage of massed practice did not show up on old/practiced words was probably because with old/practiced words, both massed and distributed groups had reached ceiling or best possible performance for their ability level (especially musical ability) on oral tone production accuracy after three intensive training sessions. The two ISI groups were equal on musical aptitude and were trained to the best possible they could reach; therefore, the advantage of massed practice was washed out. However, with new words/linguistic contexts that none of the participants had practiced before, the advantage of massed practice showed up, probably only the massed practice group was able to reach the deeper procedural knowledge or level of automatization in oral tone production. This advantage of massed practice over distributed practice on oral word naming accuracy of new words confirms that procedural knowledge is the dominant component in performing this task.

Although the main effect for RI did not turn out to be statistically significant, F (1, 59) = 3.606, p =.062, the effect size was medium, η_p^2=.058. Descriptively, the 4-week RI seems to have resulted in lower tone production accuracy than 1-week RI at each of the respective ISI levels (see Figure 12). This is probably due to the declarative component involved in completing this task. With the long 4-week RI, participants' declarative knowledge about the tones might have gone through some decay, minor probably, which might have resulted in the lower accuracy in oral tone production when facing new words, which requires a fusing process for pronouncing tones in new combinations of segments. Further research is needed to test whether this prediction holds.

Taking together the results on written picture naming, oral word naming on old words, and oral word naming on new words, this study shows that the effects of ISI and RI do differ depending on the type of knowledge/skill to be retained, i.e., purely declarative versus skill. For the retention of declarative knowledge, RI has a robust effect, the longer the RI, the more the forgetting, and the worse the retention. Spacing, or distributed practice seems to improve long-term retention of declarative knowledge; however, this ISI effect is much weaker than the RI effect. With regard to skill retention, in contrast, ISI turned out to play a significant role (at least for the retention of oral tone production skill on new words), but not RI (at least weaker). Furthermore, it was massed practice and not distributed practice that had the advantage for skill retention, which is consistent with the finding from Suzuki (2017).

Having examined the effects of ISI and RI on the two subcomponents, let us now turn to the global complex skill that is comprised of the two subcomponents. Oral picture naming is a complex skill that requires both the declarative component of meaning-word mappings and the procedural component in oral word/tone articulation. No strong hypothesis was formulated regarding the effects of ISI and RI on the retention of this complex skill, but I anticipated that performance on oral picture naming may be first constrained by the retention of the declarative component (i.e., picture-Pinyin mappings) and then by the retention of the procedural component in oral articulation. The involvement of declarative knowledge about the tones (DK2) is optional because it is not required if this knowledge is fully automatized. I therefore anticipated that Group A may outperform the other three groups in oral

picture naming accuracy, at least descriptively. This indeed turned out to be true (see Figure 13).

For retention performance in oral picture naming accuracy, a significant main effect for RI was found; longer 4-week RI resulted in much lower oral picture naming accuracy than the shorter 1-week RI. Neither a main effect for ISI, nor an interaction between ISI and RI was found. This pattern of results on oral picture naming accuracy parallels the results on written picture naming accuracy (compare Figure 9 and Figure 14). It seems that the mechanism for the retention of the subcomponent of declarative knowledge was vital for the retention of the complex skill that comprises it. It is logical that retention performance on the complex skill would be first constrained by the retention of the declarative component of this skill, because if a participant forgot the word for the picture, they would not be able to orally produce the word, even if s/he would be able to pronounce tones perfectly when given Pinyin. For the second subcomponent in oral articulation, no advantage was observed for massed practice, just as no advantage was observed for oral word naming on old/practiced words (because all words in oral picture naming were old words too). As for the interaction between ISI and RI, again, this interaction was not statistically significant either in oral picture naming accuracy. Results from pairwise comparisons showed that the effect of ISI was negligible for the short 1-week RI, F (1, 64) = 0.361, p = .550, η_p^2=.006, and seemed to be larger for the longer 4-week RI, F (1, 64) = 2.914, p = .093, η_p^2=.044 with comparatively larger effect size though still statistically nonsignificant. This pattern of results regarding the interaction between ISI and RI on *oral* picture naming parallels with the results on *written* picture naming,

in that while there is little effect of ISI on retention for the short RI, there seems to be an effect of spacing for the long RI (with small effect size on oral picture naming accuracy η_p^2=.044, and medium effect size on written picture naming η_p^2=.074). This pattern of results suggests that the effect of ISI depends on RI, which is in line with Cepeda et al.'s (2006) prediction.

As for retention performance on oral picture naming RT, there was a main effect for both RI and ISI; specifically, longer RI resulted in slower RT, but longer ISI resulted in faster RT (see Figure 15). A task component analysis situated in the context of this training study can help understand the latter effect. Performing this oral picture-naming task entails two cognitive processes: picture-word retrieval and orally producing the word. From the oral word-naming task, we knew that RT did not differ between groups. It is thus reasonable to attribute the ISI effect on RT to the first cognitive process in retrieving picture-word mappings. It is to be expected, therefore, that the distributed group (1-week ISI) may be faster in retrieving picture-word mappings than the massed group (1-day ISI) at the retention test administered weeks (1 week or 4 weeks) after the last training session, because the 1-week ISI groups practiced retrieving words from pictures on a weekly basis in the second and third training sessions, while the 1-day ISI groups only practiced to retrieve words on a daily basis during the training sessions and it was their first time trying to retrieve words after a week or 4 weeks delay on the retention test. Therefore, this advantage for distributed practice over massed on retention performance in oral picture naming RT should be attributed to the cognitive process in retrieving picture-word mappings, which the distributed practice group practiced on an ISI schedule that is on a similar

scale of the RI; this advantage should not be interpreted as that distributed practice led to increased automatization or higher level of automaticity in oral articulation (the second subcomponent). As I did not use an RT measure for the written picture naming task, it is impossible to compare the pattern of the ISI and RI effects on oral picture naming with that of written picture naming, but I would expect that the patterns would be parallel.

It seems, then that for retention of a complex skill that requires both a declarative component and a procedural component, with the declarative component being a priori for completing the task, the effects of ISI and RI on the complex task are constrained by the retention mechanism of the declarative component.

7.2 The effect of Cognitive Aptitudes on L2 learning of Mandarin tonal word production

This study also examined the roles of cognitive aptitudes (including working memory, phonological STM, declarative memory, procedural memory, and musical aptitude) at different stages of learning oral Mandarin word production and on the retention of the learned knowledge or skills. Table 55 summarizes the results on the presence or absence of statistically significant relationships with the outcome measures at different stages, i.e., at the end of the first training session, at the end of the last training session, and finally at the retention test after a delay, with effects of ISI and RI controlled.

For the beginning stage of learning, it was hypothesized that DM ability, WM, PSTM, and musical aptitude would be a significant predictor of oral picture naming accuracy at the end of the first training session. None of these predictions was borne

out, however. The learning outcome at this early stage was not predicted by WM capacity, DM ability, PSTM, or musical aptitude; instead, it was predicted by pretest performance in tone production accuracy in disyllabic words, which was then further predicted by musical aptitude. Musical aptitude did not directly predict oral picture naming accuracy at the end of TS1; instead, this relationship was mediated by oral tone production accuracy at the pretest. The other cognitive aptitudes, including DM ability, WM capacity, and PSTM, that were theorized to play a role at the initial stage of learning (such as skill acquisition theory, and the Declarative/Procedural model about L2 learning) did not turn out to be significant predictors. It is possible that the effects of these cognitive aptitudes were masked by some other unidentified factors. More research is required to explain this result.

Table 55. *Summary of results for the role of cognitive aptitudes in outcome performance at different stages*

	TS1_Post	TS3_Post		Retention Stage						
	OPicN	OPicN		WPicN	OWN_old		OWN_new		OPicN	
	Acc	Acc	RT	Acc	Acc	RT	Acc	RT	Acc	RT
CVMT	x			√					x	
WM	x	(√)		(√)			√		√	
NWR	x								x	
Music	x	√			√		√		√	
SRT*		x	x			√		x		√
						ISI*SRT				RI*SRT

Note. All were based on single-predictor/covariate analyses after controlling ISI or ISI & RI when appropriate; * after further controlling L1 word naming RT; "√" means a statistically significant relationship was found; "x" means a statistically significant relationship was not found; significance was set at $p < .05$; () were results from follow-up analyses; blank cells are those not tested.

Post-session performance at the last training session (TS3) was considered a later stage of learning in the context of this study. Musical aptitude was found to be a

significant predictor of oral picture naming accuracy as hypothesized. PM ability as measured by SRT however was not significantly related to either the accuracy measure (p=.131, η_p^2 = .035) or the RT measure (p=.155, η_p^2 = .032) of oral picture naming at the end of TS3. A follow-up analysis was then done to see whether WM still plays a role at this stage, and it was found that WM was positively related to oral picture naming accuracy in the post-session quiz of TS3.

At the retention stage, for oral picture naming performance, WM and musical aptitude also turned out to be significant predictors for oral picture naming accuracy, as hypothesized. DM ability as measured by CVMT and PSTM as measured by NWR however did not turn out to be significant predictors. For retention performance in oral picture naming RT, a statistically significant interaction between SRT and RI was found (p = .038), and the interaction between SRT and ISI was marginal (p = .086). To better interpret the interaction effects, it may help to look at the relationships between SRT and retention performance in oral picture naming in terms of both accuracy and RT measures across all four groups (see Table 56). None of the correlation coefficients were statistically significant. In Group A (daily practice with short 1-week RI), the positive relationship between SRT and oral picture naming accuracy seemed strong (r = .372), suggesting a facilitative role, but the correlation with RT was almost negligible (r = .025). In Group B (daily practice with longer 4-week RI), in contrast, the correlation of SRT with accuracy was negligible (r = .074), but it seems to be strongly correlated with RT and the correlation was positive (r = .415) (higher PM ability, slower). In Group C (weekly practice with short 1-week RI), SRT did not seem to correlate with accuracy but seemed to negatively correlated with

RT (higher PM ability, faster). In Group D, SRT did not seem to correlate with either accuracy or RT. This complex pattern is very puzzling, especially that while PM ability seems to play a facilitative role in the two short 1-week RI groups (Group A with accuracy; Group C with RT), higher PM ability seems to slow down RT in Group B, the group with daily massed practice but with long 4-week RI. It seems that with such a long retention interval, it took those with higher PM ability longer to recall and orally produce the words correctly. Note that the oral picture naming RT was based on only 34.4% of all RT data points (see 6.2.2) as 65.6% of the RT data were excluded due to high error rates. On top of that, Group B had the lowest accuracy scores across the four groups in this oral picture naming task and thus the smallest amount of RT data points for analysis. Therefore, the interaction effects should be interpreted with caution. More research is required to validate the interactions and see how the role of PM ability might be mediated by ISI or RI.

Table 56. *Correlations between SRT and retention performance in oral picture naming (Accuracy & RT) across experimental conditions*

	D5_OPicN_4com_Acc	D5_OPicN_RT_M_lg
Grp A: ISI-1d; RI-1w	.372	.025
Grp B: ISI-1d; RI-4w	.074	.415
Grp C: ISI-1w; RI-1w	.070	-.144
Grp D: ISI-1w; RI-4w	.005	-.038

Note. None of the correlations were statistically significant.

For the retention of the declarative component of word knowledge in meaning-word mappings, DM ability as measured by CVMT was found to be a significant predictor of written picture naming accuracy, as hypothesized. Those with higher DM ability remembered the Pinyin of the words (spelling plus tone marks for

each syllable) better. In addition, working memory was also found a significant predictor of written picture naming accuracy on the retention test; those with higher working memory capacity also remembered the Pinyin of the words better after a RI. CVMT and WM independently predicted retention performance in written picture naming accuracy. When both CVMT and WM were put into the same model, however, the effect of CVMT turned out to be nonsignificant while WM was still a robust predictor. Thus, it seems that WM capacity plays a facilitative role independently from DM ability on the retention of declarative knowledge. It should also be noted however that the WM measures have higher reliability (.719 for Shapebuilder and .799 for Ospan) than the CVMT measures (.673 for CVMT-acquisition and .660 for CVMT-delayed), which could be why the findings for CVMT turned out to be less robust.

For the retention of oral tone production skill on old/practiced words (the other subcomponent), musical aptitude was found a significant predictor of oral tone production accuracy as hypothesized; those with higher musical aptitude retained higher oral tone production accuracy. As for the effect of SRT on oral word naming RT on practiced words, an interaction between SRT and ISI was found: for the 1-week ISI, the relationship between PM ability as measured by SRT and oral word naming RT on practiced words was negative (R= - .155), suggesting that for the weekly practice groups, participants with higher PM ability tended to respond faster in oral word naming on practiced words; for the 1-day ISI, in contrast, the relationship between PM ability and oral word naming RT was positive (R= .197), suggesting that for the daily practice groups, participants with higher PM ability

tended to respond slower in oral word naming on practiced words on the retention test. See Table 57 for the correlations between SRT and retention performance in oral word naming on old words in terms of both accuracy and RT across the four groups. For accuracy, it seems that PM ability may play a facilitative role in Groups A and C, the two groups with the short 1-week RI. When the RI was longer, 4 weeks, PM ability plays different roles, i.e., speeding up RT in Group D (distributed practice), but slowing down RT in Group B (massed practice). Again, as the RT analysis was based on limited RT data points (50.1%), this interaction effect should be interpreted with caution. More research is required to understand how PM ability might interact with practice schedules and retention intervals.

Table 57. *Correlations between SRT and retention performance in oral word naming on old words (Accuracy & RT) across experimental conditions*

	D5_OWN_old_Tone_Acc	D5_OWN_old_RT_M_lg
Grp A: ISI-1d; RI-1w	.241	.092
Grp B: ISI-1d; RI-4w	-.032	.292
Grp C: ISI-1w; RI-1w	.277	-.109
Grp D: ISI-1w; RI-4w	.053	-.235

Note. None of the correlations were statistically significant.

Finally, for the retention of oral tone production skill on new words, WM and musical aptitude turned out to be significant predictors for oral tone production accuracy as hypothesized. WM capacity seems to have played a facilitative role in the fusing process of pronouncing tones with new combinations of segmental sounds. Those with higher musical aptitude were able to pronounce the tones in new words more accurately. While PM ability was hypothesized to play a facilitative role in the

oral word naming RT measure, this prediction was not borne out: SRT was not found to be significantly related to oral word naming RT on new words.

The most robust predictors (aptitudes) of the learning outcomes for Mandarin word production in the context of this study seem to be musical aptitude and WM capacity. Musical aptitude was a robust predictor of all accuracy outcome measures involving oral production, including oral picture naming accuracy at the end of TS3 and at the retention test, and oral tone accuracy in oral word naming on old words and on new words on the retention test; for oral picture naming accuracy at the end of TS1, it was not directly predicted by musical aptitude, but it was predicted by oral tone production accuracy at the pretest that was further predicted by musical aptitude. As long as oral production of Mandarin tonal words was involved in the outcome tests, musical aptitude played a facilitative role for production accuracy across the board at different stages. This finding is consistent with the positive relationship found between musical ability and monosyllabic Mandarin tone-word perception and production performance at the end of training found in Li and DeKeyser (in press). The results of the present study further extended this relationship between musical aptitude and Mandarin word learning to *disyllabic* word production and to learning outcome at different stages including the learning stages and the retention stage.

Related to musical aptitude, it would be interesting to see whether musical training experience is a predictor of outcome performance in tonal word learning. Among the 68 participants included in final analysis, 57 participants had not had formal musical training experience, while 11 reported having formal musical training experience (ten had had private musical lessons, and one self-taught in two

instruments). Musical training was therefore coded at these two levels and used as a categorical variable. An ANOVA was first run to see whether the group with formal musical training experience (n=11) had significantly higher musical scores than the group with no formal musical training experience (n=57), and it turned out to be true, F (1, 66) = 8.375, p = .005. Table 58 presents the distribution of participants with or without formal musical training experience across the four experimental groups; the 11 participants with formal training were sparsely distributed across the groups.

Table 58. *Distribution of participants with or without formal musical training across the four experimental groups*

		Experimental Groups				
		ISI-1d; RI-1w	ISI-1d; RI-4w	ISI-1w; RI-1w	ISI-1w; RI-4w	Total
Musical Training	No	13	15	15	14	57
	Yes	5	3	1	2	11
Total		18	18	16	16	68

Further analyses were conducted to see whether musical training experience was a predictor of outcome performance in oral Mandarin tonal word production. A series of linear regressions were conducted with musical training as a predictor, ISI and RI included as appropriate, and the five oral production accuracy measures as dependent variables. Table 59 presents the results from such analyses. As expected, musical training as a categorical variable was a much weaker predictor of outcome performance. It only turned out to be statistically significant on oral word naming accuracy on old words on the retention test. On the other hand, it is interesting to see that this categorical variable, whether participants had formal musical training or not,

seems to be a promising predictor of outcome performance, at least in oral tone production accuracy in the oral word naming task. Further research needed to see whether formal musical training (with vs. without) is a robust predictor of L2 learning of tonal word production.

Table 59. *Musical training experience as a predictor on outcome performance (when controlling for ISI or ISI & RI when appropriate)*

Outcome Measures	*B*	*SEB*	*Standardized Beta*	*t*	*p*
TS1_Pre_OWN_Tone_Acc_sqrt	.039	.023	.207	1.720	.090
TS1_Post_OPicN_4com_Acc	-.002	.021	-.015	-.119	.905
TS3_Post_OPicN_4com_Acc	.016	.015	.128	1.028	.308
D5_OWN_old_Tone_Acc	.073	.031	.287	2.370	.021
D5_OWN_new_Tone_Acc	.052	.031	.207	1.715	.091
D5_OPicN_4com_Acc	.022	.022	.113	1.000	.321

WM capacity was also a robust predictor of L2 learning outcome of Mandarin word production at different stages (both the learning stage and the retention stage) across a range of tasks, including oral picture naming accuracy at the end of TS3 and on the retention test, and on retention performance in written picture naming accuracy and oral word naming accuracy with new words. The fact that WM still plays a role in oral picture naming accuracy at the end of the last training session suggests that participants still use controlled processing in oral picture naming at the end of TS3. It is worth exploring whether this controlled processing was for the cognitive retrieval of picture-word mappings or for oral word/tone articulation, or for both but to

different degrees. This finding of a robust facilitative role of complex WM in L2 learning of Mandarin word production is consistent with previous findings in Martin and Ellis (2012) and Kapa and Colombo (2014) for the role of the executive function component of WM in L2 word learning, especially in L2 learning of oral word production.

DM ability as measured by CVMT was hypothesized to play a role in oral picture naming accuracy at the beginning stage of learning (based on Ullman's the Declarative/Procedural model, and the findings from Morgan-Short et al., 2014) and on retention performance (due to the declarative component required to complete the task), but the predictions were not confirmed. DM ability was indeed found to facilitate retention of declarative knowledge as measured by written picture naming accuracy. The fact that CVMT predicted written picture naming accuracy but not oral picture naming accuracy on the retention test seems to suggest that oral picture naming relies less on declarative memory than written picture naming. On other hand, CVMT as a measure of DM ability is relatively new, and its reliability was relatively low; more work is needed to increase the reliability of the CVMT task.

PSTM was hypothesized to play a facilitative role in oral picture naming at the beginning stage of learning and on retention; however, these predictions were not borne out. This finding seems to be inconsistent with the previous results from Wei (2015), who found that PSTM independently predicted Chinese spoken word learning (including 1-syllable, 2-syllable, and 3-syllable words) in native English-speaking children (4^{th} grade, 10-11 years old), and from Bowles et al. (2016) who found PSTM (as measured by a nonword span task) a significant predictor of tonal word learning

success in sound-word identification accuracy. More research is required to see what factors might have contributed to this inconsistency.

The role of PM ability seems to be most complicated in the automatization of L2 oral Mandarin tonal word production. It was hypothesized that PM ability would start to play a role at the later stage of learning and on the retention, based on the Declarative/Procedural model (Ullman, 2001, 2004, 2015) and the research findings from Morgan-Short et al. (2014) and Hamrick (2015). PM ability as measured by SRT, however, did not turn out to play a significant role at the end of the TS3. This result seems to be inconsistent with Morgan-Short et al. (2014) who found that syntactic performance at the late stage of learning positively correlated with procedural learning ability; however these two studies differ in so many ways that it is hard to pinpoint what has contributed to the different findings. Three factors stood out which might explain the inconsistency. First, there were four training sessions in Morgan-Short et al. (2014) whereas there were only three training sessions in the present study. More practice may be required for the learning mechanism to shift from relying on declarative memory to relying on procedural memory. Second, Morgan-Short et al. (2014) used implicit training conditions whereas the present study used explicit training. It is possible that the implicit training conditions are more conducive to learning relying on procedural memory. Third, in Morgan-Short et al. (2014), procedural memory ability was measured by a Weather Prediction Task (probabilistic), and a Tower of London Task (cognitive skill learning); in the present study, PM ability was measured by using SRT. The SRT task used in this study is arguably a more implicit measure.

For the oral picture naming performance at the end of TS3 in this study, the CV analysis of RT suggests that automatization in terms of restructuring or stability had occurred by the end of the TS3 (for details see 6.2.2). In other words, though the CV analysis suggests that automatization has occurred, this automatized learning outcome was not predicted by PM ability. Given the limited number of training sessions, it is possible that most participants have reached the incipient stage of automatization, and the knowledge they have developed was "explicit automatized knowledge" (Suzuki & DeKeyser, 2017), which is probably why WM was still a significant predictor of oral picture naming accuracy at this stage, and SRT was not. Another reason why SRT did not turn out to be a significant predictor of outcome performance at this later stage could be the relatively low reliability of SRT. Note that the correlations between SRT and outcome performance in accuracy at the end of TS3 (r = .144 for the massed group; r = .238 for the distributed group) were stronger than the correlation between them at the end of TS1 (r = .036) (see Appendix L & M), suggesting that PM ability started to play a stronger role at the later stage compared to the initial stage.

After a retention interval, i.e., given more time, PM ability as measured by SRT was found to interact with RI for the retention performance in oral picture naming RT, and interact with ISI for the retention performance in oral word naming RT on old words. Thus, there is indeed a change over time regarding the role of PM ability on L2 learning of oral tonal word production. For retention performance in oral picture naming RT, though the relationship between SRT and the retention RT was not statistically significant in separate RI groups, the direction of the relationship

was the opposite, i.e., the relationship was negative (thus facilitative) for the short-term 1-week RI, but the relationship was positive for the long-term 4-week RI (higher PM ability, slower RT). For retention performance in oral word naming RT on old words, the relationship between PM ability and the retention RT was negative (thus facilitative) for the longer 1-week ISI, but was positive (higher PM ability, slower RT) for the shorter 1-day ISI. Suzuki (2017b) found that procedural learning ability measured by the Tower of London task was only significantly associated with RT in the 3.3-day ISI group (facilitative role), but not in the 7-day ISI group. It seems that the facilitative role of SRT is bounded by a time range, not too short (such as 1-day ISI), not too long (e.g., 4-week RI). The exact learning mechanisms under different practice distribution conditions still await to be discovered.

Finally, L1 word naming RT was used as a covariate to control individual differences in basic cognitive processing speed, because some people are just faster and others slower even in L1 processing. According to Ackerman's theory regarding the three stages of skill acquisition, performance in the third/autonomous stage is determined by psychomotor ability, or "psychophysical limitations of the human subject" (Ackerman, 1988, p. 291). As English NSs have fully automatized their L1 skills, variation in L1 word naming RT can be a good indicator of psychomotor ability in word production. Then, a strong correlation with L1 word naming RT can be an indication of a high level of automaticity. Among the four RT outcome measures, L1 word naming RT turned out to be a significant covariate for three of them, i.e., the three RT measures from tasks with old/practiced words (oral picture naming at the end of TS3 and on the retention test, and oral word naming on old

words on the retention test), indicating high level of automaticity on producing the practiced words was achieved by the end of TS3 and retained in both oral picture naming and oral word naming after a retention interval. The only RT outcome measure that L1 word naming RT was not a significant covariate for was the oral word naming RT on new words, indicating the involvement of cognitive processing when facing new words in this task. In addition, L1 word naming RT was found to interact with RI on retention performance in oral word naming RT of old words. It was found that L1 word naming RT was more strongly correlated with the outcome RT for 1-week RI (R =.685) than for 4-week RI (R =.214), suggesting that the level of automaticity in oral word naming skills was better retained at the short 1-week retention interval than at the long 4-week retention interval. The results of this study highlight the importance of including L1 word naming RT as a covariate to control individual differences in basic cognitive processing.

Chapter 8: Conclusions

The present dissertation research adds to the current body of literature on temporal distribution of practice suggesting that the effects of ISI and RI differ depending on the type of knowledge/skill to be retained, purely declarative versus skill. For the retention of declarative knowledge, RI has a robust effect: the longer the RI, the more the forgetting, and the worse the retention. Spacing, or distributed practice seems to improve long-term retention of declarative knowledge; however, this ISI effect is much weaker. With regard to procedural knowledge retention, in contrast, ISI seems to play an important role, but not RI, and it was massed practice that had an advantage over distributed practice for skill retention. For the retention of a complex skill that entails both a declarative component and a procedural component, with the declarative component being a prerequisite for performing the skill, the effects of ISI and RI on the complex task are constrained by the retention mechanisms of the declarative component.

This study provides empirical evidence that while declarative knowledge is vulnerable to memory decay (longer RI, worse retention), procedural knowledge or highly automatized skill is much more robust and much less susceptible to decay. The empirical evidence from this study lends strong support to the skill retention theory (Kim et al., 2013). By examining the effects of ISI and RI on the subcomponents of a complex skill that differ in nature (declarative vs. skill), and finding that the effects do differ depending on type of knowledge/skill to be retained, this study highlights the importance of practicing the subcomponents of a complex skill at different and

concerted schedules for most effective proceduralization and optimal retention of each of the subcomponents or subskills.

For skill acquisition across the training sessions, the advantage for massed practice was found at the beginning of the training sessions, but this advantage did not hold through to the end of the training sessions. These results suggest that the advantage of massed practice can or at least can appear to be overridden by effective training.

This study also scrutinized the role of cognitive aptitudes in L2 learning of Mandarin tonal word production. Musical aptitude turned out to be a significant predictor of oral word/tone production accuracy across the board at different stages. Working memory also played a facilitative role at both the learning stage and for retention in oral word production from conceptualization to articulation, in tone production when facing new words, and in recalling declarative knowledge of word forms. Phonological short-term memory did not turn out to be a significant predictor. Declarative memory ability was a significant predictor of retention performance in a declarative task. As for the role of procedural memory ability, it seems to play a stronger role at the later stage of learning compared to the initial stage; in addition, its role on skill retention in terms of response time seems to be mediated by intersession interval (ISI) or retention interval (RI). There is indeed a change over time regarding the role of PM ability on L2 learning of oral tonal word production. More research is required to examine the role of PM ability at different stages of learning. More theoretically driven research testing how cognitive aptitudes may interact with ISI or RI would also likely be fruitful.

Future L2 research should also systematically investigate the effects of ISI and RI on the learning and retention of complex skills, such as the automatization (receptive or productive) of grammatical rules, being phonological, morphological, or syntactic, by using task/skill component analysis, to explore the optimal practice distributions for the procedualization and retention of the subcomponents as well as the global complex skill. Another interesting question worth of future investigation is how temporal distribution of practice interacts with stages of skill acquisition; answers to that question will contribute not only to theory but also to practice. Future L2 research should also systematically manipulate different levels of ISI and RI to figure out optimal practice schedules for different types of knowledge or skill, or complex skills. Meanwhile, this study would benefit from replication to check generalizability.

Appendix A. Participant Background Questionnaire

1. Age: _____

2. Gender: _____Male _____Female

3. Major: _____

4. Student status: _____Undergraduate _____Graduate student ____Not a student

5. Native language: _____

6. a) Do you speak another language? _____Yes _____No

b) Did you learn another language? _____Yes _____No

c) If yes to any of the above questions, please list the other language(s) and provide relevant information for each.

Language	Proficiency on a scale of 1 to 10 (1=minimal; 10= near-native)	At what age did you start to learn it?	In which context(s) did you learn it?

7. a) Did you have any exposure to Chinese, Cantonese, Thai or Vietnamese?

____Yes ____No

b) If yes, in which context(s), and for how long? _____

8. a) Did you have any formal training in any musical instrument (including voice)?

____Yes ____No

b) If yes, please list the instruments and provide relevant information for each.

Instrument	At what age did you start to learn it?	For how long you took lessons in it?

9. Do you have hearing loss? _____

10. Do you have speech impairment? _____

Appendix B. Target Words

yīnghuā  cherry blossom	wūpó  witch	bānmǎ  zebra	wāndòu  pea
máoyī  sweater	yínháng  bank	píngguǒ  apple	tóufà  hair
kǔguā  bitter melon	kǒnglóng  dinosaur	fěnbǐ  chalk	lǐngdài  tie
dàngāo  cake	dìtú  map	dàogǔ  paddy	pùbù  waterfall
lǎohǔ  tiger	gǎnlǎn  olive	lǐpǐn  gift	gǎngkǒu  harbor

Appendix C. Items for the Generalization Test

List 1	List 2	List 3	List 4
gūmā	fēnfāng	hēimāo	gōngkāi
kōnglíng	pīnbó	tuōpín	sēnlín
gāndǎn	tūdǐng	dāibǎn	gēnběn
bīpò	tōukàn	tōngtòu	wēndù
bíyīn	huábīng	tíkū	fángdōng
wúdí	púfú	yáolán	nóngmín
lípǔ	wánměi	lóuyǐ	fánnǎo
láogù	fénmù	táigàng	wénhuà
dǒupō	tǔbō	huǒguō	kǎoyā
fǎguó	wǔtái	mǎnyíng	lǎngdú
yǐndǎo	dǐdǎng	hǎodǎi	suǒyǒu
gǎnwù	lǒngluò	nǎilào	kěndìng
kùdōu	bànpāi	lòukōng	fàngsōng
kòngfá	yìngpán	tàihú	gòngtóng
hùbǔ	guòmǐn	dàoyǐng	dàmǐ
màoyì	lànmàn	wèibì	làngfèi

Appendix D. EI Sheet Introducing Mandarin Chinese

Introduction to Mandarin Chinese

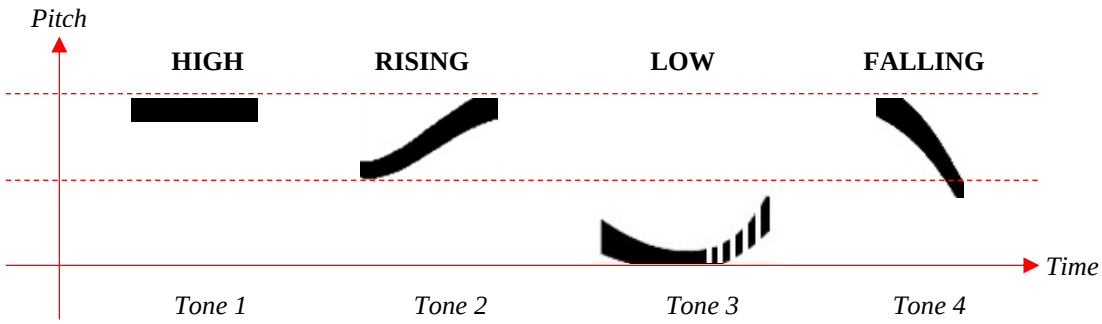
In this study, you are going to learn some Chinese words, in particular, Chinese words in Mandarin, which is the standard spoken language in Mainland China. Your goal in this study is to learn twenty Mandarin words consisting of two syllables each. Before we delve into that, I'd first like to give you an introduction to the sound system of Mandarin Chinese.

The first thing you may need to know about Mandarin Chinese is that Mandarin is a *tone* language. Each word in Mandarin carries a tone. Tones are differences in pitch that can change the *meaning* of a word. Mandarin distinguishes between four different tones, or pitch patterns (e.g., bā, bá, bǎ, bà). That is, *ba* can have four different meanings depending on the tone (see the table below): bā means "eight", bá means "to pull out", bǎ means "target", and bà means "father". In other words, to say a Mandarin word requires that you use the correct tone for that word.

bā	bá	bǎ	bà
8 **eight**	 **to pull out**	 **target**	 **father**

Mandarin tones are marked by symbols above the vowels when Chinese is written with the English alphabet. This way of writing Chinese is called **Pinyin**. The first row in the above table gives the Pinyin for the four words.

Now let's get to how to produce the tones. The following diagram gives the visual representations of the tones. The vertical represents the range of your speaking pitch. This range is probably somewhat greater than the range you normally use in English. The top may seem slightly too high and the bottom slightly too low.



The first tone is called the *high* tone (e.g., bā). It starts high, stays high and doesn't change while you're producing it. Its pitch is near the top of your speaking range and it is level, not going up or down. The second tone is called the *rising* tone (e.g., bá). It starts in the middle of your range, and rises right away, to the top of your range. The third tone is called the *low* tone. It starts low, then dips to the bottom of your range (e.g., bă). It sometimes scoops up a bit (e.g., bă) (still within the lower register of your voice); therefore it has the scooping tone mark symbol ˇ. Note however that the rising tail, which is depicted in dashed curve above, *only* occurs when it is spoken in isolation or at utterance-final position for emphasis. It's most often just a *low* tone in other contexts. The fourth tone is a *falling* tone (e.g., bà). It starts high and then drops sharply. It's a sudden drop, and this tone is the shortest. To recap, we have a *high* tone, a *rising* tone, a *low* tone, and a *falling* tone.

Now let's practice hearing and producing these tones in some single syllables, before moving on to focus on the learning of the words with two syllables.

Appendix E. Preprogramed Lists for Tone Practice in Monosyllables

A. For Tone Identification Practice

Round 1	mo1, bai1, tan1, hong1, mo2, bai2, tan2, hong2, mo3, bai3, tan3, hong3, mo4, bai4, tan4, hong4
Round 2	bai3, hong3, mo3, tan3, mo2, tan2, bai2, hong2, mo1, hong1, tan1, bai1, hong4, mo4, tan4, bai4
Round 3	mo1, mo2, mo3, mo4, bai1, bai3, bai4, bai2, tan4, tan1, tan2, tan3, hong2, hong4, hong3, hong1

B. For Tone Production Practice

Round 1	mo1, bai1, tan1, hong1, mo2, bai2, tan2, hong2, mo3, bai3, tan3, hong3, mo4, bai4, tan4, hong4
Round 2	mo1, bai1, tan1, hong1, mo2, bai2, tan2, hong2, mo3, bai3, tan3, hong3, mo4, bai4, tan4, hong4
Round 3	mo1, mo2, mo3, mo4, bai1, bai2, bai3, bai4, tan1, tan2, tan3, tan4, hong1, hong2, hong3, hong4

Appendix F. Sheet for Tone Identification Practice in monosyllables

mō	mó	mǒ	mò
bāi	bái	bǎi	bài
tān	tán	tǎn	tàn
hōng	hóng	hǒng	hòng

Appendix G. EI Sheet Introducing Tone Changes in Disyllabic Words

Tonal changes in two-syllable words

As the majority of Mandarin words have two syllables, e.g., nǐhǎo, which means "hello", this study is going to focus on such words.

Now that you are able to identify and produce the four tones in single syllables, let's see what happens when we combine tones in two-syllable words. Tone 1, Tone 2, and Tone 4, i.e., the high, rising, and falling tones, stay basically the same in two-syllable words. Tone 3, or the low tone, in two-syllable words, is most often *just low*, without the rising tail. It changes, however, when it is followed by another Tone 3. When a low tone comes before another low tone, the first low tone changes to a *rising* tone, i.e., Tone 2, for example, nǐhǎo "hello". Before the low tone of hǎo, the low tone of nǐ becomes a rising tone "ní". Other than in this situation, Tone 3 is just low in disyllabic words.

Tone 3 in two-syllable words

a) **T3 becomes a Rising tone/T2**, when it is *followed* by another T3

níhǎo

"hello"

b) **Otherwise, T3 is just a low tone**, in any other contexts, i.e., (i) when it is followed by T1, T2, and T4, and (ii) when it is on the last syllable of a word

Appendix H. Sheet for Picture-Pinyin Mapping Practice

yīnghuā	wūpó	bānmǎ	wāndòu
máoyī	yínháng	píngguǒ	tóufà
kǔguā	kǒnglóng	fěnbǐ	lǐngdài
dàngāo	dìtú	dàogǔ	pùbù
lǎohǔ	gǎnlǎn	lǐpǐn	gǎngkǒu

Appendix I. Stimuli for the L1 Word Naming Task

museum
hero
onion
arrow
candy
dragon
baseball
city
album
walnut
lotus
sticker
menu
donkey
eagle
bunny
maple
mountain
oven
diamond
blanket
forest
salad
guitar
kitchen
feather
pizza
castle
ocean
dolphin

Appendix J. Scoring Instruction to the 2nd rater

A. Background of the oral word production tasks

- See either a picture or the Pinyin of a word, and participants were asked to either say the Chinese word for the picture they see or to read the Pinyin aloud
  - Oral picture naming
  - Oral word naming
- The audio recordings of their word production are the targets for scoring

B. Scoring criteria

- For each item, participants were asked to orally produce a *two-syllable* word, e.g., wūpó for *witch*.
- Each two-syllable word is given 4 points in total, 1 point for each of the following 4 components, i.e.,
  - 1st syllable *segments*
  - 1st syllable *tone*
  - 2nd syllable *segments*
  - 2nd syllable *tone*
- *Tone* is scored dichotomously,
  - i.e., either as correct (√, 1 point) or incorrect (×, 0); no partial points
  - *How to judge correctness of tones*, especially of tones pronounced by learners? More details later
- *Segments* can receive partial credit (√` 0.5)
  - The segments of a syllable is broken down into two components, i.e., *ping*
    - *i.* initial: p (0.5)
    - *ii.* final: ing (0.5)
  - when both the initial and the final for a syllable are pronounced correct, mark √ (1 point); when only the initial or only the final is pronounced correct, mark √` (0.5 point); if neither is pronounced correct, then mark × (0 point)
  - underline the segment that is incorrectly pronounced, if credit is taken off
- If there's no oral production for a given item, mark -------- above the word, and note down "missing"

- See a scoring sheet for example. √, ×, √`

EXAMPLE SCORING SHEET

C. Format of data entering for each trial in each task per participant

		1st syllable		2nd syllable	
TrialNo.	TargetWords	Segments	Tone	Segments	Tone
1	wūpó	1	0	1	0
2	píngguǒ	1	0	1	0
3	kǔguā	1	1	1	0
4	pùbù	1	0	1	0
5	lǎohǔ	1	0	1	0
6	yínháng	1	0	1	1
7	yīnghuā	1	0	0.5	0
8	fěnbǐ	1	0	1	0
9	lǐpǐn	1	0	0.5	0
10	dìtú	1	0	1	1
11	wāndòu	1	0	1	0
12	máoyī	0.5	0	1	0
13	gǎnlǎn	1	0	0.5	0
14	lǐngdài	1	1	1	0
15	dàogǔ	1	0	1	0
16	bānmǎ	1	0	0.5	0
17	tóufà	1	0	1	0
18	gǎngkǒu	1	0	1	0
19	kǒnglóng	1	1	1	1
20	dàngāo	1	0	0.5	0

D. Tone Scoring Rubric

- How were they taught?
    - See the instruction sheets participants received

- Scoring criteria for correctness of tones
    - Tone 1: high + stays high/flat/consistent;
        - if not high enough, ×;
        - if not flat, ×;
        - if sounds like rising, ×;
        - if sounds like falling, ×.

    - Tone 2: start low and then go high;
        - if sounds just high, and can not hear rising, ×.

    - Tone 3 (LOW): start low, then go even lower, except when it's followed by another T3, then the first T3 is pronounced as a rising tone (or T2); T3 is low-to-lower in any other contexts, i.e., word-initial position NOT followed by T3 and word-final position.
        - If there's a rising tail at the word-initial position, ×（恐龙，苦瓜，领带）
        - T3 at the word initial position, if pronounced as rising or there's a rising tail, ×

    - Tone 4: start high, then a quick drop;
        - if sounds like Tone 1, ×;
        - if rise first and then fall, ×;
        - if low-falling (i.e., does not start high enough), ×

- If non-target-like, then mark ×

How to treat false starts and self-corrections?

- score their first try

E. Segments Scoring Rubric

No.	Pinyin	Notes for scoring
1	yīnghuā	ying: back nasal, if pronounced as front nasal "in", then X; hua: h if pronounced as w, X
2	wūpó	wu: u with lips more forward; if pronunced as the Eng ʊ as in *put*, X po: o not əʊ
3	bānmǎ	ban: b, if pronounced as the Eng "b", okay
4	wāndòu	wan: an, not æn dou: d, if pronounced as the Eng "d", okay; ou: NOT uo
5	máoyī	ao:
6	yínháng	ang: NOT æŋ
7	píngguǒ	ping: NOT pin guo: g, if pronounced as the Eng "g", okay; uo NOT əu, au, etc.
8	tóufà	tou: t, if pronounced as the Eng "t", okay; ou, NOT au, or uo
9	kǔguā	u:
10	kǒnglóng	ong: NOT ɑŋ; ɒŋ, ɔŋ accepted
11	fěnbǐ	en: ən, NOT ɛn
12	lǐngdài	dai: d, if pronounced as the Eng "d", okay
13	dàngāo	dan: d, if pronounced as the Eng "d", okay; an NOT æn gao: g, if pronounced as the Eng "g", okay
14	dìtú	di: d, if pronounced as the Eng "d", okay tu: t, if pronounced as the Eng "t", okay
15	dàogǔ	dao: d, if pronounced as the Eng "d", okay gu, u NOT the Eng ʊ
16	pùbù	u, not the Eng ʊ
17	lǎohǔ	u, not the Eng ʊ
18	gǎnlǎn	gan: g, if pronounced as the Eng "g", okay; an, NOT æn
19	lǐpǐn	li: i NOT ɪ pin: p is p^h, NOT p unaspirated; in NOT ɪn, but "in"
20	gǎngkǒu	gang: g, if pronounced as the Eng "g", okay; ang NOT æŋ kou: k is k^h, not k; ou, not au, uo, etc.

Appendix K. Stimuli for the Non-Word Repetition Task

Length of three	gadge, norb, kern pem, tudge, bon chorg, parn, jit narg, derb, bup
Length of four	doob, marn, tep, cham jat, nerch, teeb, gop geed, nam, dorch, charn barp, doog, keb, dorl
Length of five	nug, mab, gerp, teeg, darch teep, nart, ped, bordge, goot mun, gerk, dort, chim, narb, jert, noog, pab, turg, larm
Length of six	padge, meb, kerm, barg, pim, dordge garm, jerg, neeb, chull, borp, narp neeg, peb, garn, tam, gerb, kig jarm, norg, putch, terdge, darp, bick

Appendix L. Correlations between the five aptitude construct scores and the pre- and post-session quiz performance on TS1

N=68

		TS1_Pre_OWN_Tone_Acc_sqrt	CVMT_ZAcq Delayed	WM_ZShapeBOspan	NWR_Z	Music_Z2PitchSTM1PTM	SRT_Z
TS1_Pre_OWN	*r*		**0.027**	**0.170**	**-0.120**	**0.291***	**0.049**
_Tone_Acc_sqrt	*p*		0.828	0.165	0.332	0.016	0.692
TS1_Post_OPicN	*r*	**0.300***	**0.185**	**0.143**	**-0.017**	**0.017**	**0.036**
_4com_Acc	*p*	0.013	0.130	0.243	0.892	0.891	0.769

* Correlation is significant at the 0.05 level (2-tailed).

Appendix M. Correlations between the five aptitude construct scores and the post-session quiz performance on TS3 across the two ISI groups

ISI=1DAY		CVMT_Z AcqDelayed	WM_Z ShapeB Ospan	NWR_Z	Music_Z 2PitchSTM 1PTM	SRT_Z
TS3_Post _OPicN_ 4com_**Acc**	*r*	**0.204**	**0.187**	**-0.102**	**0.288**	**0.144**
	p	0.233	0.276	0.554	0.088	0.403
	n	36	36	36	36	36
TS3_Post _OPicN_ **RT**_M_lg	*r*	**-0.128**	**0.001**	**-0.007**	**0.13**	**0.213**
	p	0.464	0.994	0.968	0.457	0.219
	n	35	35	35	35	35

ISI=1WEEK		CVMT_Z AcqDelayed	WM_Z ShapeB Ospan	NWR_Z	Music_Z 2PitchSTM 1PTM	SRT_Z
TS3_Post _OPicN_ 4com_**Acc**	*r*	**0.279**	**0.418***	**0.072**	**0.274**	**0.238**
	p	0.122	0.017	0.695	0.129	0.191
	n	32	32	32	32	32
TS3_Post _OPicN_ **RT**_M_lg	*r*	**-0.026**	**0.341**	**-0.236**	**0.113**	**0.055**
	p	0.888	0.061	0.201	0.544	0.768
	n	31	31	31	31	31

* Correlation is significant at the 0.05 level (2-tailed).

Appendix N. Correlations between the five aptitude construct scores and the outcome measures (Accuracy & RT) on the retention test across groups

Accuracy Outcome Measures

Group = ISI-1d; RI-1w (n=18)		CVMT_Z AcqDelayed	WM_Z ShapeBOspan	NWR_Z	Music_Z 2PitchSTM1PTM	SRT_Z
D5_OPicN_4com_Acc	*r*	**0.529***	**-0.023**	**-0.202**	**0.297**	**0.372**
	p	0.024	0.927	0.423	0.231	0.129
D5_WPicN_4comp_Acc	*r*	**0.673****	**0.373**	**-0.142**	**0.059**	**0.329**
	p	0.002	0.128	0.573	0.815	0.182
D5_OWN_old_Tone_Acc	*r*	**0.351**	**0.289**	**-0.298**	**0.19**	**0.241**
	p	0.153	0.245	0.23	0.45	0.336
D5_OWN_new_Tone_Acc	*r*	**0.125**	**0.248**	**-0.322**	**0.372**	**0.166**
	p	0.622	0.322	0.193	0.128	0.511

Group = ISI-1d; RI-4w (n=18)		CVMT_Z AcqDelayed	WM_Z ShapeBOspan	NWR_Z	Music_Z 2PitchSTM1PTM	SRT_Z
D5_OPicN_4com_Acc	*r*	**-0.137**	**0.454**	**0.109**	**0.167**	**0.074**
	p	0.589	0.059	0.666	0.507	0.772
D5_WPicN_4comp_Acc	*r*	**-0.001**	**0.428**	**-0.097**	**-0.137**	**-0.098**
	p	0.998	0.077	0.701	0.589	0.699
D5_OWN_old_Tone_Acc	*r*	**-0.177**	**-0.029**	**-0.133**	**0.418**	**-0.032**
	p	0.481	0.909	0.6	0.084	0.901
D5_OWN_new_Tone_Acc	*r*	**0.237**	**0.345**	**-0.277**	**0.141**	**-0.08**
	p	0.344	0.161	0.265	0.576	0.751

Group = ISI-1w; RI-1w (n=16)		CVMT_Z AcqDelayed	WM_Z ShapeBOspan	NWR_Z	Music_Z 2PitchSTM1PTM	SRT_Z
D5_OPicN_4com_Acc	*r*	**0.114**	**0.228**	**0.036**	**0.396**	**0.070**
	p	0.673	0.396	0.896	0.129	0.798
D5_WPicN_4comp_Acc	*r*	**0.335**	**0.307**	**-0.048**	**0.444**	**-0.051**
	p	0.205	0.247	0.86	0.085	0.851
D5_OWN_old_Tone_Acc	*r*	**-0.019**	**0.199**	**-0.017**	**0.232**	**0.277**
	p	0.944	0.459	0.95	0.387	0.298

D5_OWN_new_Tone_Acc	*r*	**0.064**	**0.447**	**0.158**	**0.138**	**0.448**
	p	0.814	0.083	0.558	0.61	0.082

Group = ISI-1w; RI-4w (n=16)		CVMT_Z AcqDelayed	WM_Z ShapeBOspan	NWR_Z	Music_Z 2PitchSTM 1PTM	SRT_Z
D5_OPicN_4com_Acc	*r*	**0.595***	**0.292**	**0.046**	**0.259**	**0.005**
	p	0.015	0.273	0.867	0.332	0.987
D5_WPicN_4comp_Acc	*r*	**0.293**	**0.329**	**-0.016**	**0.257**	**-0.018**
	p	0.27	0.213	0.953	0.338	0.948
D5_OWN_old_Tone_Acc	*r*	**0.33**	**0.127**	**0.057**	**0.331**	**0.053**
	p	0.212	0.64	0.833	0.211	0.845
D5_OWN_new_Tone_Acc	*r*	**0.153**	**0.524***	**0.120**	**0.579***	**0.325**
	p	0.573	0.037	0.657	0.019	0.219

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

RT Outcome Measures

Group = ISI-1d; RI-1w		EngOWN_RT_M_lg	CVMT_ZAcq Delayed	WM_Z ShapeBOspan	NWR_Z	Music_Z 2PitchSTM 1PTM	SRT_Z
D5_OPicN_RT_M_lg	*r*	**0.534**	**-0.668****	**-0.386**	**-0.169**	**-0.034**	**0.025**
	p	0.022	0.002	0.113	0.504	0.893	0.922
	n	18	18	18	18	18	18
D5_OWN_old_RT_M_lg	*r*	**0.784**	**-0.107**	**-0.155**	**-0.176**	**-0.035**	**0.092**
	p	0	0.682	0.553	0.5	0.895	0.724
	n	17	17	17	17	17	17
D5_OWN_new_RT_M_lg	*r*	**0.402**	**0.002**	**-0.446**	**-0.32**	**0.175**	**0.454**
	p	0.098	0.995	0.064	0.195	0.487	0.058
	n	18	18	18	18	18	18

Group = ISI-1d; RI-4w		EngOWN_RT_M_lg	CVMT_ZAcq Delayed	WM_Z ShapeBOspan	NWR_Z	Music_Z 2PitchSTM 1PTM	SRT_Z
D5_OPicN_RT_M_lg	*r*	**0.309**	**0.158**	**-0.079**	**-0.15**	**0.08**	**0.415**
	p	0.244	0.56	0.771	0.579	0.769	0.11
	n	16	16	16	16	16	16
D5_OWN_old_RT_M_lg	*r*	**0.242**	**-0.044**	**0.024**	**0.07**	**0.162**	**0.292**
	p	0.367	0.872	0.929	0.797	0.549	0.272
	n	16	16	16	16	16	16

D5_OWN_new_RT_M_lg	*r*	**0.041**	**-0.008**	**0.297**	**-0.246**	**-0.488***	**0.141**
	p	0.875	0.977	0.247	0.341	0.047	0.589
	n	17	17	17	17	17	17

Group = ISI-1w; RI-1w		EngOWN_RT_M_lg	CVMT_ZAcq Delayed	WM_Z ShapeB Ospan	NWR_Z	Music_Z 2PitchSTM 1PTM	SRT_Z
D5_OPicN_RT_M_lg	*r*	**0.367**	**-0.343**	**-0.24**	**-0.127**	**-0.013**	**-0.144**
	p	0.24	0.275	0.452	0.694	0.968	0.656
	n	12	12	12	12	12	12
D5_OWN_old_RT_M_lg	*r*	**0.524***	**-0.448**	**-0.3**	**-0.161**	**-0.003**	**-0.109**
	p	0.045	0.094	0.277	0.566	0.99	0.7
	n	15	15	15	15	15	15
D5_OWN_new_RT_M_lg	*r*	**0.303**	**-0.192**	**-0.089**	**-0.129**	**0.348**	**-0.181**
	p	0.254	0.476	0.744	0.634	0.187	0.502
	n	16	16	16	16	16	16

Group = ISI-1w; RI-4w		EngOWN_RT_M_lg	CVMT_ZAcq Delayed	WM_Z ShapeB Ospan	NWR_Z	Music_Z 2PitchSTM 1PTM	SRT_Z
D5_OPicN_RT_M_lg	*r*	**0.29**	**0.264**	**0.389**	**-0.116**	**-0.083**	**-0.038**
	p	0.294	0.341	0.152	0.68	0.77	0.894
	n	15	15	15	15	15	15
D5_OWN_old_RT_M_lg	*r*	**0.078**	**0.067**	**0.163**	**0.043**	**-0.075**	**-0.235**
	p	0.774	0.806	0.546	0.873	0.782	0.38
	n	16	16	16	16	16	16
D5_OWN_new_RT_M_lg	*r*	**-0.086**	**0.471**	**0.28**	**-0.193**	**0.555***	**0.052**
	p	0.752	0.065	0.293	0.473	0.026	0.847
	n	16	16	16	16	16	16

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Bibliography

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288.

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, *6*(4), 259.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–406. https://doi.org/10.1037/0033-295X.89.4.369

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, N.J.: L. Erlbaum Associates.

Anderson, J. R. (2000). *Learning and memory: an integrated approach* (2nd ed.). New York: Wiley. Retrieved from Table of contents http://catdir.loc.gov/catdir/toc/onix01/99032553.html

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*(4), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Antoniou, M., & Wong, P. C. M. (2016). Varying irrelevant phonetic features hinders learning of the feature being trained. *The Journal of the Acoustical Society of America*, *139*(1), 271–278.

Atkins, P. W. B., & Baddeley, A. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics*, *19*(4), 537–552. https://doi.org/10.1017/S0142716400010353

Atkins, S. M., Sprenger, A. M., Colflesh, G. J. H., Briner, T. L., Buchanan, J. B., Chavis, S. E., … Dougherty, M. R. (2014). Measuring working memory is all fun and games: A four-dimensional spatial game predicts cognitive task performance. *Experimental Psychology*, *61*(6), 417–438. https://doi.org/10.1027/1618-3169/a000262

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. (2003). Working Memory and Language: An Overview. *Journal of Communication Disorders*, *36*(3), 189–208.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

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. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742108604521

Baddeley, A., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language, 27*(5), 586–595. https://doi.org/10.1016/0749-596X(88)90028-9

Beaunieux, H., Hubert, V., Witkowski, T., Pitel, A.-L., Rossi, S., Danion, J.-M., … Eustache, F. (2006). Which processes are involved in cognitive procedural learning? *Memory, 14*(5), 521–539. https://doi.org/10.1080/09658210500477766

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*(4), 635–650.

Bloom, K. C., & Shuell, T. J. (1981). Effects of Massed and Distributed Practice on the Learning and Retention of Second-Language Vocabulary. *The Journal of Educational Research*, *74*(4), 245–248.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.08. Retrieved from http://www.praat.org/

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch Ability As an Aptitude for Tone Learning. *Language Learning, 66*(4), 774-808. https://doi.org/10.1111/lang.12159

Carpenter, H. S. (2008). *A behavioral and electrophysiological investigation of different aptitudes for L2 grammar in learners equated for proficiency level* (Unpublished Ph.D. Dissertation). Georgetown University. Retrieved from https://repository.library.georgetown.edu/handle/10822/558127

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771.

Carron, A. V. (1969). Performance and Learning in a Discrete Motor Task under Massed Vs. Distributed Practice. *Research Quarterly.American Association for Health, Physical Education and Recreation*, *40*(3), 481–489. https://doi.org/10.1080/10671188.1969.10614866

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236–246.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention. *Psychological Science*, *19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual Variability in Cue-Weighting and Lexical Tone Learning. *Journal of the Acoustical Society of America*, *128*(1), 456–465.

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *The Journal of the Acoustical Society of America*, *138*(6), 3703–3716.

Chao, Y. R. (1930). A system of tone-letters. *Le Maître Phonétique*, *45*(1), 24–27.

Chao, Y. R. (1948). *Mandarin primer: An intensive course in spoken Chinese*. Cambridge: Harvard University Press.

Chen, M. Y. (2000). *Tone Sandhi: Patterns Across Chinese Dialects* (Vol. 92). Cambridge: Cambridge University Press.

Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary language. *Developmental Psychology*, *32*(5), 867–873. https://doi.org/10.1037/0012-1649.32.5.867

Collins, L., Halter, R. H., Lightbown, P. M., & Spada, N. (1999). Time and the Distribution of Time in L2 Instruction. *TESOL Quarterly*, *33*(4), 655–680. https://doi.org/10.2307/3587881

Collins, L., & White, J. (2011). An Intensive Look at Intensity and Language Learning. *TESOL Quarterly*, *45*(1), 106–133. https://doi.org/10.5054/tq.2011.240858

Conti-Ramsden, G., Ullman, M. T., & Lum, J. A. G. (2015). The relation between receptive grammar and procedural, declarative, and working memory in specific language impairment. *Frontiers in Psychology*, *6*(Article 1090). https://doi.org/10.3389/fpsyg.2015.01090

Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *The Journal of the Acoustical Society of America*, *131*(6), 4756–4769.

Cooper, A., & Wang, Y. (2013). Effects of tone training on Cantonese tone-word learning. *The Journal of the Acoustical Society of America*, *134*(2), 133–139.

Dail, T. K., & Christina, R. W. (2004). Distribution of practice and metacognition in learning and long-term retention of a discrete motor task. *Research Quarterly for Exercise and Sport*, *75*(2), 148–155.

DeKeyser, R. M. (2007a). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, New Jersey: Lawrence Erlbaum.

DeKeyser, R. M. (2007b). Study abroad as foreign language practice. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 208–226). Cambridge: Cambridge University Press.

DeKeyser, R. M. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning*, *62*(Suppl. 2), 189–200. https://doi.org/10.1111/j.1467-9922.2012.00712.x

DeKeyser, R. M. (2015). Skill Acquisition Theory. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction* (pp. 94–112). New York, NY: Routledge.

DeKeyser, R. M., & Criado, R. (2012). Automatization, Skill Acquisition, and Practice in Second Language Acquisition. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley-Blackwell. Retrieved from http://dx.doi.org/10.1002/9781405198431.wbeal0067

DeKeyser, R. M., & Prieto Botana, G. (2015). The Effectiveness of Processing Instruction in L2 Grammar Acquisition: A Narrative Review. *Applied Linguistics*, *36*(3), 290–305. https://doi.org/10.1093/applin/amu071

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805.

Duanmu, S. (2007). *The Phonology of Standard Chinese*. New York: Oxford University Press.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. New York: Dover Publications (Original work published 1885).

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, *40*(1), 83–107.

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 243–285). New York: Academic Press.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, and Computers*, *35*(Journal Article), 116–124.

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, *29*(3), 463–487. https://doi.org/10.1017/S0142716408080211

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The*

Quarterly Journal of Experimental Psychology Section A, *54*(1), 1–30. https://doi.org/10.1080/02724980042000002

Gluckman, M., Vlach, H. A., & Sandhofer, C. M. (2014). Spacing Simultaneously Promotes Multiple Forms of Learning in Children's Science Curriculum. *Applied Cognitive Psychology*, *28*(2), 266–273. https://doi.org/10.1002/acp.2997

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In P. Rebuschat (Ed.), *Implicit and Explicit Learning of Languages* (pp. 443–482). Amsterdam / Philadelphia: John Benjamins.

Hamrick, P. (2015). Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences*, *44*, 9–15. https://doi.org/10.1016/j.lindif.2015.10.003

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 47–91). New York: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742108604648

Hubert, V., Beaunieux, H., Chételat, G., Platel, H., Landeau, B., Danion, J.-M., … Francis Eustache. (2007). The dynamic network subserving the three phases of cognitive procedural learning. *Human Brain Mapping*, *28*(12), 1415–1429. https://doi.org/10.1002/hbm.20354

Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics*, *30*(2), 225–249. https://doi.org/10.1017/S0142716409090109

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649–667. https://doi.org/10.1016/S0022-5371(78)90393-6

Kang, S. H. K., & Pashler, H. (2012). Learning Painting Styles: Spacing is Advantageous when it Promotes Discriminative Contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. https://doi.org/10.1002/acp.1801

Kapa, L. L., & Colombo, J. (2014). Executive function predicts artificial language learning. *Journal of Memory and Language*, *76*, 237–252. https://doi.org/10.1016/j.jml.2014.07.004

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction*, *36*, 38–45. https://doi.org/10.1016/j.learninstruc.2014.11.001

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340. https://doi.org/10.1016/j.cognition.2010.05.011

Kaushanskaya, M. (2012). Cognitive mechanisms of word learning in bilingual and monolingual adults: The role of phonological memory. *Bilingualism:*

Language and Cognition, *15*(3), 470–489. https://doi.org/10.1017/S1366728911000472

Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science*, *14*(1), 22–37.

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. https://doi.org/10.1017/S1366728908003416

Kornell, N., & Bjork, R. A. (2008). Learning Concepts and Categories: Is Spacing the "Enemy of Induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Küpper-Tetzel, C., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*(1), 37–47. https://doi.org/10.1080/09658211.2011.631550

Küpper-Tetzel, C., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, *42*(3), 373–388. https://doi.org/10.1007/s11251-013-9285-2

Lapkin, S., Hart, D., & Harley, B. (1998). Case study of compact core French models: Attitudes and achievement. In S. Lapkin (Ed.), *French second language education in Canada: Empirical studies* (pp. 3–30). Toronto: University of Toronto Press.

Leather, J. (1990). Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In J. Leather & A. James (Eds.), *New Sounds 90: Proceedings of the Amsterdam Symposium on the Acquisition of Second Language Speech*. Amsterdam: University of Amsterdam.

Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport*, *59*(4), 277–287.

Lee, T. D., & Genovese, E. D. (1989). Distribution of practice in motor skill acquisition: Different effects for discrete and continuous tasks. *Research Quarterly for Exercise and Sport*, *60*(1), 59–65.

Leeser, M. J. (2007). Learner-Based Factors in L2 Reading Comprehension and Processing Grammatical Form: Topic Familiarity and Working Memory. *Language Learning*, *57*(2), 229–270. https://doi.org/10.1111/j.1467-9922.2007.00408.x

Li, M., & DeKeyser, R. (in press). Perception practice, production practice, and musical ability in Mandarin tone-word learning. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263116000358

Li, S. (2010). The effectiveness of corrective feedback in SLA: a meta-analysis. *Language Learning*, *60*(2), 309–365.

Lightbown, P. M., & Spada, N. (1994). An Innovative Program for Primary ESL Students in Quebec. *TESOL Quarterly*, *28*(3), 563–579. https://doi.org/10.2307/3587308

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. https://doi.org/10.1111/lang.12011

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), 861–83.

Linck, J. A., & Weiss, D. J. (2011). Working memory predicts the acquisition of explicit L2 knowledge. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 101–113). Washington, DC: Georgetown University Press.

Lum, J. A. G., Conti-Ramsden, G., Page, D., & Ullman, M. T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex*, *48*(9), 1138–1154. https://doi.org/10.1016/j.cortex.2011.06.001

Lyster, R., & Saito, K. (2010). Oral Feedback in Classroom SLA. *Studies in Second Language Acquisition*, *32*(2), 265–302.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed second language acquisition* (Vols. 1–Book, Section, pp. 181–209). Amsterdam: Benjamins.

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. https://doi.org/10.1017/S0272263112000125

Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory*, *13*(3–4), 422–429.

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System*, *42*, 412–428. https://doi.org/10.1016/j.system.2014.01.014

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K. A., Carpenter, H., & Wong, P. C. M. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, *17*(1), 56–72. https://doi.org/10.1017/S1366728912000715

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R. V., & Vul, E. (2009). Predicting the Optimal Spacing of Study: A Multiscale Context Model of Memory. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1321–1329). Curran Associates, Inc.

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning. *Studies in Second Language Acquisition*, *37*(4), 677–711. https://doi.org/10.1017/S0272263114000825

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (Vol. 1, pp. 1–55). Hillsdale, NJ: Erlbaum.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-analysis. *Language Learning*, *50*(3), 417–528.

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. https://doi.org/10.1017/S027226310707043X

Opitz, B., Schneiders, J. A., Krick, C. M., & Mecklinger, A. (2014). Selective transfer of visual working memory training on Chinese character learning. *Neuropsychologia*, *53*, 1–11. https://doi.org/10.1016/j.neuropsychologia.2013.10.017

Paik, J., & Ritter, F. E. (2016). Evaluating a range of learning schedules: hybrid training schedules may be as good as or better than distributed practice for some tasks. *Ergonomics*, *59*(2), 276–290. https://doi.org/10.1080/00140139.2015.1067332

Panchuk, D., Spittle, M., Johnston, N., & Spittle, S. (2013). Effect of practice distribution and experience on the performance and retention of a discrete sport skill. *Perceptual & Motor Skills: Learning & Memory*, *116*(3), 750–760.

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language*, *30*(3), 331–347. https://doi.org/10.1016/0749-596X(91)90040-Q

Pavlik, P. I., & Anderson, J. R. (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. *Cognitive Science*, *29*(4), 559–586. https://doi.org/10.1207/s15516709cog0000_14

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a Novel Phonological Contrast Depends on Interactions between Individual Differences and Training Paradigm Design. *Journal of the Acoustical Society of America*, *130*(1), 461–472.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. *Cognitive Science*, *27*(3), 431–452. https://doi.org/10.1016/S0364-0213(03)00007-7

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude, and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfi eld, and Hernstadt (1991). In P. Robinson (Ed.), *Individual*

differences in instructed language learning (pp. 181–209). Amsterdam: Benjamins.

Roehr, K. (2013). Aptitude-treatment interaction (ATI) research. In P. Robinson (Ed.), *The Routledge Encyclopedia of Second Language Acquisition* (pp. 31–35). New York, NY: Routledge.

Rogers, J. (2015). Learning Second Language Syntax Under Massed and Distributed Conditions. *TESOL Quarterly*, *49*(4), 857–866. https://doi.org/10.1002/tesq.252

Rohrer, D. (2015). Student Instruction Should Be Distributed Over Long Time Periods. *Educational Psychology Review*, *27*(4), 635–643. https://doi.org/10.1007/s10648-015-9332-4

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*(4), 183–186.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*(9), 1209–1224.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498. https://doi.org/10.1007/s11251-007-9015-8

Sanz, C., Lin, H.-J., Lado, B., Stafford, C. A., & Bowden, H. W. (2014). One Size Fits All? Learning Conditions and Working Memory Capacity in Ab Initio Language Development. *Applied Linguistics*, amu058. https://doi.org/10.1093/applin/amu058

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, *19*(1), 28–42. https://doi.org/10.1177/1362168814541726

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*(1), 107–122. https://doi.org/10.1002/acp.1066

Serafini, E. J., & Sanz, C. (2015). Evidence for the decreasing impact of cognitive ability on second language development as proficiency increases. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263115000327

Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning*, *61*(1), 117–145.

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System*, *35*(3), 305–321. https://doi.org/10.1016/j.system.2007.02.001

Shea, C. H., Lai, Q., Black, C., & Park, J.-H. (2000). Spacing practice sessions across days benefits the learning of motor skills. *Human Movement Science*, *19*(5), 737–760. https://doi.org/10.1016/S0167-9457(00)00021-X

Shintani, N. (2015). The Effectiveness of Processing Instruction and Production-based Instruction on L2 Grammar Acquisition: A Meta-Analysis. *Applied Linguistics*, *36*(3), 306–325. https://doi.org/10.1093/applin/amu067

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-Based Versus Production-Based Grammar Instruction: A Meta-Analysis of Comparative Studies. *Language Learning*, *63*(2), 296–329. https://doi.org/10.1111/lang.12001

Simmons, A. L. (2012). Distributed Practice and Procedural Memory Consolidation in Musicians' Skill Learning. *Journal of Research in Music Education*, *59*(4), 357–368. https://doi.org/10.1177/0022429411424798

Slevc, L. R., & Miyake, A. (2006). Individual Differences in Second-Language Proficiency: Does Musical Ability Matter? *Psychological Science*, *17*(8), 675–681.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767.

Spada, N., & Tomita, Y. (2010). Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis. *Language Learning*, *60*(2), 263–308.

Staddon, J. E. R., Chelaru, I. M., & Higa, J. J. (2002). Habituation, memory and the brain: the dynamics of interval timing. *Behavioural Processes*, *57*(2–3), 71–88. https://doi.org/10.1016/S0376-6357(02)00006-2

Suzuki, Y. (2017a). The Optimal Distribution of Practice for the Acquisition of L2 Morphology: A Conceptual Replication and Extension. *Language Learning*, *67*(3), 512–545. https://doi.org/10.1111/lang.12236

Suzuki, Y. (2017b). The role of procedural learning ability in automatization of L2 morphology under different learning schedules: An exploratory study. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263117000249

Suzuki, Y., & DeKeyser, R. (2015). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*. https://doi.org/10.1177/1362168815617334

Suzuki, Y., & DeKeyser, R. (2016). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics*. https://doi.org/10.1017/S0142716416000084

Suzuki, Y., & DeKeyser, R. (2017). The Interface of Explicit and Implicit Knowledge in a Second Language: Insights From Individual Differences in Cognitive Aptitudes. *Language Learning*. https://doi.org/10.1111/lang.12241

Tagarelli, K. M., Ruiz, S., Vega, J. L. M., & Rebuschat, P. (2016). Variability in second language learning: The roles of individual differences, learning conditions, and linguistic complexity. *Studies in Second Language Acquisition*, *38*(2), 293–316. https://doi.org/10.1017/S0272263116000036

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, *15*(5), 529–536. https://doi.org/10.1016/0022-5371(76)90047-5

Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *The Psychology of Learning and Motivation*, *60*, 113–189.

Trahan, D. E., & Larrabee, G. J. (1988). *Continuous Visual Memory Test*. Odessa, FL: Assessment Resources.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Bilingualism: Language and Cognition*, *4*(1), 105–122.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition*, *92*(1–2), 231–270.

Ullman, M. T. (2015). The Declarative/ Procedural Model: A neurobiologically motivated theory of first and second language. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction* (Vol. 2nd, pp. 135–158). New York, NY: Routledge.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498–505.

Vatz, K., Tare, M., Jackson, S. R., & Doughty, C. J. (2013). Aptitude-treatment interaction studies in second language acquisition: Findings and methodology. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 273–292). Amsterdam: Benjamins.

Verhagen, J., & Leseman, P. (2016). How do verbal short-term memory and working memory relate to the acquisition of vocabulary and grammar? A comparison between first and second language learners. *Journal of Experimental Child Psychology*, *141*, 65–82. https://doi.org/10.1016/j.jecp.2015.06.015

Verhagen, J., Leseman, P., & Messer, M. (2015). Phonological Memory and the Acquisition of Grammar in Child L2 Learners. *Language Learning*, *65*(2), 417–448. https://doi.org/10.1111/lang.12101

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: the spacing effect in children's acquisition and generalization of science concepts. *Child Development*, *83*(4), 1137–1144. https://doi.org/10.1111/j.1467-8624.2012.01781.x

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167. https://doi.org/10.1016/j.cognition.2008.07.013

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: an investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*(5), 750–763. https://doi.org/10.3758/s13421-010-0063-y

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and Perceptual Evaluation of Mandarin Tone Productions before and after Perceptual Training. *The Journal of the Acoustical Society of America*, *113*(2), 1033–1043.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American Listeners to Perceive Mandarin Tones. *Journal of the Acoustical Society of America (JAS)*, *106*(6), 3649–3658.

Wei, J. (2015). Prediction on English-speaking children's Chinese spoken word learning: Contributions of phonological short-term memory. *Arizona Working Papers in SLA & Teaching*, *22*, 101–129.

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.

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(4), 565–585.

Woodcock, R. W. (1997). The Woodcock-Johnson Tests of Cognitive Ability—Revised. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 230–246). New York, NY, US: Guilford Press.

Wright, C. (2013). An Investigation of Working Memory Effects on Oral Grammatical Accuracy and Fluency in Producing Questions in English. *TESOL Quarterly*, *47*(2), 352–374. https://doi.org/10.1002/tesq.68

Yang, C. (2016). *The Acquisition of L2 Mandarin Prosody: From experimental studies to pedagogical practice* (Vol. 1). Amsterdam: John Benjamins Publishing Company. Retrieved from http://www.jbe-platform.com/content/books/9789027267634

Yazdani, M. A., & Zebrowski, E. (2006). Spaced reinforcement: An effective approach to enhance the achievement in plane geometry. *Journal of Mathematical Sciences and Mathematics Education*, *1*(3). Retrieved from http://ww.w.msme.us/2006-1-8.pdf

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.

Zhang, H. (2016). The effect of theoretical assumptions on pedagogical methods: a case study of second language Chinese tones. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.12132

Zulkiply, N., & Burt, J. S. (2012). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*(1), 16–27. https://doi.org/10.3758/s13421-012-0238-9

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction*, *22*(3), 215–221. https://doi.org/10.1016/j.learninstruc.2011.11.002