WORKING MEMORY, PRESENTATION FORMATS, AND ATTENTION: AN EYE-TRACKING STUDY ON LEARNING L2 CHINESE CHARACTERS IN A COMPUTER-ASSISTED SELF-STUDY ENVIRONMENT

By

Xuehong He

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies—Doctor of Philosophy

2020

ABSTRACT

WORKING MEMORY, PRESENTATION FORMATS, AND ATTENTION:
AN EYE-TRACKING STUDY ON LEARNING L2 CHINESE CHARACTERS
IN A COMPUTER-ASSISTED SELF-STUDY ENVIRONMENT

By

Xuehong He

Drawing on the recent framework of internal and external attention in cognitive science (Chun et al., 2011), the current study explored how learner internal and external factors, namely, working memory capacities and presentation formats affected learner attention and learning outcome. Sixty-nine English native speakers studied 30 two-character Chinese words in three different presentation formats, namely, horizontal, vertical, and adjacent, within a computer-assisted self-study context. Their learning gains were measured with a bilingual vocabulary test that adopted recognition and recall tasks to assess different mappings between form and meaning. Learners' eye movements when viewing the characters, pinyin, and English meaning of the Chinese words were recorded during the learning process. Two attention indices were employed: fixation durations and fixation counts. Working memory capacities were assessed with a storage, an inhibition, a shifting, and an updating tasks based on Miyake et al.'s (2000) framework. Mixed effects modeling and repeated-measures ANOVA, as well as descriptive statistics and bivariate correlations were conducted for data analysis.

Results showed that compared with the horizontal and vertical formats, the adjacent format generally led to better learning outcome and promoted attention to the characters, when factors including vocabulary test formats and L2 Chinese proficiency were taken into consideration. Working memory capacities were also generally found as a significant predictor of learner attention and learning outcome. In addition, learning outcome was predicted by learner attention. These results were discussed in terms of theoretical and pedagogical implications.

ACKNOWLEDGEMENTS

The completion of my dissertation would not have been possible without the strong support from my dissertation committee and I would like to express my sincere gratitude to Dr. Shawn Loewen (Chair), Dr. Aline Godfroid, Dr. Susan Gass, and Dr. Xiaoshi Li for their guidance throughout the years. Particularly, I want to thank Dr. Godfroid for giving detailed and insightful feedback on the research design and the data analysis of my dissertation. My greatest thanks go to my advisor and the Chair of my dissertation committee, Dr. Loewen, who has always supported me through all sorts of situations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Research on second language (L2) instruction has long been devoted to exploring effective manipulation of learner attention, with the idea that paying attention to L2 input will facilitate L2 development (Robinson, Mackey, Gass, & Schmidt, 2012; although see a different view by VanPatten, 2017). The significance of attention in L2 instruction mirrors the prominence of attention as a psycholinguistic construct in L2 research (Robinson et al., 2012; Schmidt, 2001). Attention has permeated discussions of theoretical and instructional issues fundamental in second language acquisition (SLA), and has been explored in different settings to account for diverse SLA phenomena (for review see Robinson et al., 2012).

The latest L2 studies on attention manipulation (e.g., Indrarathne & Kormos, 2016, 2017; Issa & Morgan-Short, 2019; Issa, Morgan-Short, Villegas, & Raney, 2015) have followed a taxonomy of internal attention and external attention proposed in cognitive science. Given the ubiquity of attention in cognitive science (Chun, Golomb, & Turk-Browne, 2011), recent overviews have proposed an organizing framework of the internal and external taxonomy to advance research on attention (e.g., Chun et al., 2011; Dixon, Fox, & Christoff, 2014; Lieberman, 2007). According to Chun et al. (2011), attention is not a unitary construct but a property of multiple mechanisms. Due to limited capacity, attentional mechanisms select and modulate the most relevant information for processing, and face the challenge to sustain vigilance on the information (Chun et al., 2011). Based on the source of information, *external* attention involves selecting and modulating information through the senses to the external world, while *internal* attention concerns selection and modulation of information generated in the mind (Chun et al., 2011). Within Chun et al.'s (2011) framework, attentional control can be driven by stimuli with exogenous, bottom-up processes, and can be directed by goals with endogenous,

top-down processes. Working memory is at the interface of internal and external attention, as it constrains the processing of external stimuli on the one hand, and enables information processing without external support on the other hand (Chun et al., 2011; see also Kiyonaga & Egner, 2013).

Following the taxonomy of internal and external attention, the current study aims to contribute to the line of research on attention and L2 learning, and to explore two important factors of attentional control: working memory and L2 input (see also Indrarathne & Kormos, 2017). Specifically, I will investigate how working memory and presentation formats affect learning L2 Chinese characters at the beginning level in a computer-assisted self-study environment.

Language teachers have long been seeking to apply emerging technology to improving L2 learning (Chun, Smith, & Kern, 2016; Reinders & Stockwell, 2017; Warschauer, 1996). On the other hand, the pervasiveness of technology in daily life has also placed a demand on language teachers to optimize the design and delivery of teaching and learning materials with available technology (Chun et al., 2016; Heift & Chapelle, 2012). With the advent of computers and internet, online learning has gained momentum in higher education in the United States, with over 5.8 million students taking at least one online course as reported in 2015 (Allen, Seaman, Poulin, & Straut, 2016). In special times such as the COVID-19 pandemic period, online learning has remained almost the only form of language instructions across the world (see Foreign Language Annals 2020 summer issue). Moving traditional face-to-face courses partially or even fully online has been cost-effective for both universities (Allen & Seaman, 2010) and students (Clinefelter & Aslanian, 2016), and many college programs have explored technology-supported teaching and learning of foreign languages, such as Spanish virtual (e.g., Russell, 2012) and English flipped classrooms (e.g., Hung, 2015). The positive effects of technology on L2 learning

have been confirmed in a recent meta-analysis (Grgurović, Chapelle, & Shelley, 2013), and researchers proposed that L2 studies should pay more attention to the process of computer-assisted language learning rather than merely comparing learning outcomes with and without technology (Reinders & Stockwell, 2017). Additionally, research on the design features of technology-supported language learning can provide guidance for materials developers, teachers, and learners for more efficient L2 instruction (Heift & Chapelle, 2012). Specifically, process-oriented measures, such as eye-tracking, can reveal the actual process and individual variability of learner interaction with the design features in technology-supported language learning, and can complement the sometimes unclear outcome data from pretest and posttest, so that the efficacy of the design features can be effectively evaluated (Chun et al., 2016).

The current study aims to explore learning L2 Chinese characters in a computer-assisted self-study environment so as to provide implications for future development of online vocabulary learning modules. Presentation formats as a task feature and working memory as an individual difference will be investigated with the eye-tracking technology to tap into the learning process and outcomes.

CHAPTER 2 LITERATURE REVIEW

2.1 Attention and L2 Learning

In SLA theory, one important distinction was made by Corder (1967): while the target language available to learners provides *input*, input may not be internalized by leaners as *intake*. To account for L2 learning during the initial process of converting input to intake, L2 researchers have made various proposals to explain the underlying attentional mechanisms (Leow, Grey, Marijuan, & Moorman, 2014; Robinson et al., 2012). One influential proposal is Schmidt's Noticing Hypothesis (1990, 1995, 2001), which has been promoting research on attention in L2 learning (Godfroid & Schmidtke, 2013; Robinson et al., 2012). In Schmidt's latest review (2012), he defined *noticing* as "conscious registration of attended specific instances of language" (p. 32), and hypothesized that "input does not become intake for language learning unless it is noticed, that is, consciously registered" (p. 27). The concept of noticing has been said to be of hybrid nature, involving attention and awareness (Godfroid, Boers, & Housen, 2013; Godfroid & Schmidtke, 2013; Indrarathne & Kormos, 2016; Robinson et al., 2012). The fact that the hybridity of noticing has caused both theoretical and methodological problems, and that noticing is uncommon in the literature of psychology and cognitive science (Godfroid et al., 2013; Indrarathne & Kormos, 2016) has led researchers to propose operationalizing noticing at two levels: attention and awareness (Godfroid et al., 2013; Godfroid & Schmidtke, 2013; Robinson et al., 2012). Awareness refers to "the subjective, contentful 'feel' of experience that can be reported to others, to varying extents" (Robinson et al., 2012, p. 247). While awareness is viewed as dichotomous, namely, being either aware or unaware, attention is viewed as continuous, as attention can be in various amount (Godfroid et al., 2013)

Different online measures have been proposed to tap attention and awareness during the

process of L2 input exposure (Godfroid et al., 2013; Leow et al., 2014). One established online measure of awareness is the think-aloud protocol, which asks learners to produce verbal reports during processing (Bowles, 2010; Godfroid & Schmidtke, 2013; Godfroid & Spino, 2015; Leow et al., 2014). Recently, eye-tracking has been well received by L2 researchers as a valid and robust measure of online processing (Dussias, 2010; Frenck-Mestre, 2005; Godfroid, Winke, & Conklin, 2020; Roberts & Siyanova-Chanturia, 2013; Winke, Godfroid, & Gass, 2013), and has been recommended for examining attention (Godfroid et al., 2013; Godfroid & Schmidtke, 2013; Godfroid & Uggen, 2013; Godfroid et al., 2020; Issa et al., 2015; Indrarathne & Kormos, 2016, 2017; Leow et al., 2014; Loewen & Inceoglu, 2016; Robinson et al., 2012; Schmidt, 2012; Winke, 2013b; Winke, Gass, & Sydorenko, 2013). As Robinson et al. (2012) pointed out, supplementing verbal reports with physiological measures such as eye-tracking is a worthwhile attempt to uncover the roles of attention and awareness in L2 learning.

Although the role of awareness continues to stimulate debates on core issues in SLA, such as implicit and explicit learning (Robinson et al., 2012), attention is generally acknowledged to be essential for L2 learning (e.g., Godfroid et al., 2013; Indrarathne & Kormos, 2016). The significance of attention has been shared in instructed L2 research, in the idea that paying attention to input can facilitate learning (Robinson et al., 2012). Proposals of raising learner attention to L2 input include input enhancement (Sharwood Smith, 1981, 1991) and focus on form (Long, 1991, 1996). Along with the continuing interests in attention and input, recent studies have started to explore attention in relation to individual differences, such as motivation (e.g., Issa et al., 2015) and working memory (e.g., Indrarathne & Kormos, 2017; Mackey, Philp, Egi, Fujii, & Tatsumi, 2002). Fundamental in L2 learning, attention may serve as "a pivotal point" for connecting learner-internal (e.g., motivation, aptitude, and prior knowledge) and

learner-external factors (e.g., input, context, and task) during L2 acquisition (Schmidt, 2001, 2012, p. 44). In the following, I will review one internal factor, working memory, and one external factor, L2 input.

2.2 Working Memory and Attention

Working memory as a psychological construct has gained popularity in recent L2 research (Juffs & Harrington, 2011) and has assumed importance in SLA theory development (Linck, Osthus, Koeth, & Bunting, 2014). With its origin in cognitive science, working memory can make prolific contributions to SLA research by shedding light on not only L2 aptitude, but also cognitive processes of L2 learning (Williams, 2015). Emphasizing an important role of attention, working memory is thus highly relevant to L2 theories and instruction (Williams, 2015). In the upcoming section, I will first discuss working memory models and theories in cognitive science and their relevance to SLA, then introduce the operationalization and measures of working memory capacities, and end with discussions of L2 empirical studies on working memory and attention.

2.2.1 Working Memory in Cognitive Science and SLA

In the latest overview on the definition of working memory, Cowan (2017) pointed out that there are actually as many as *nine* different definitions in use, which are sometimes implied and sometimes clearly stated (see also Miyake & Shah, 1999). This proliferation may have resulted from the diverse measures and theoretical orientations towards working memory, and has created confusion and controversies among researchers (Cowan, 2017). It is beyond the current scope to examine all definitions, but I will focus on the three most widely adopted

approaches: Baddeley's Multicomponent Model (Baddeley, 1986, 2000; Baddeley & Hitch, 1974), Cowan's Embedded-Processes Model (1995, 1999, 2005), and Engle and colleagues' Executive-Attention Theory (Engle, Tuholski, Laughlin, & Conway, 1999; Engle & Kane, 2004). These approaches are more often discussed in SLA research (Linck et al., 2014; see Baddeley, 2015; Cowan, 2015 and Bunting & Engle, 2015 for their speculations about working memory and SLA). Finally, I will introduce Miyake and colleagues' framework of the executive component in working memory (Miyake & Friedman, 2012; Miyake, Friedman, Emerson, Witzki, & Howerter, 2000).

First introduced by Miller, Galanter, and Pribram (1960), the term *working memory* became dominant in cognitive psychology after Baddeley and Hitch's (1974) seminal Multicomponent Model (Baddeley, 2007; Cowan, 2008). Baddeley and Hitch proposed that working memory has a dual function of not only storing but also manipulating information, and consists of two domain-specific storage systems, that is, phonological loop and visuo-spatial sketchpad, and an attentional control system, namely, central executive (see also Baddeley, 1986). This Multicomponent Model was later finalized by the addition of a fourth component, episodic buffer, which is a multidimensional storage system for combining information from phonological loop, visuo-spatial sketchpad, and long-term memory (Baddeley, 2000).

As Baddeley (2012) admitted, he focused more on the storage components, particularly the phonological loop, and less on the executive component of working memory, while Cowan's major interest was the executive component. In Cowan's Embedded-Processes Model (1995, 1999, 2005), some information in the activated portion of long-term memory enters the focus of attention, whose contents are highly accessible for immediate use and under the control of central executive processes. Cowan (2008) saw working memory as consisting of storage and

attentional control systems, but he reserved the storage system for further detailed exploration rather than specifying the subsystems as Baddeley did. While Baddeley's model is structure-oriented, Cowan's model is process-oriented (Linck et al., 2014), but actually they share most views on working memory and differ mainly on emphases and terminologies (Baddeley, 2012).

With similar interest in the executive component, Engle and colleagues approached working memory from the perspective of individual differences. Whereas working memory as a *construct* may be universal to all, people do differ in their working memory *capacities*. Although Engle and colleagues based their Executive-Attention Theory (Engle et al., 1999; Engle & Kane, 2004) on Cowan's (1995, 1999, 2005) model, they disagreed with Cowan's speculation that both storage and executive systems contribute to different working memory capacities, and attributed individual differences to attentional control abilities. Subscribing to the functional importance of dual-tasking of working memory, Engle and colleagues conducted multiple complex span tasks that required online storage and manipulation of information (see Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005 for methodological review), and found substantial correlations between task scores and higher-order cognitive abilities, including general fluid intelligence (Engle et al., 1999) and language comprehension (Engle, 2001). Comparing with the often less substantial correlations between higher-order cognitive abilities and simple span tasks, Engle and Kane (2004) explained that while the simple span tasks tapped only the storage component of working memory, the complex span tasks demanded the executive component to allocate attentional resources effectively for both storage and processing, and that the different abilities in executive control caused different task performances. The findings of substantial correlations between complex span tasks and higher-order cognitive abilities were highly influential (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001), and in fact led Baddeley to

add the episodic buffer to his model (Baddeley, 2012).

Given the high relevance of working memory to cognition, L2 researchers have been working on connecting L2 acquisition with working memory (Williams, 2012, 2015). Attempts have been made to incorporate working memory research in psychology into developing L2 models of memory and attention (e.g., Robinson, 2003; Wen, 2015). However, as Williams (2012) pointed out, despite the seemingly diverse approaches to working memory, they differed mainly in emphases rather than overall conception, echoing Baddeley's (2012) view that Cowan's, Engle's, and his own approaches to working memory share overall similarities. Currently, most researchers have accepted working memory as a multicomponent system, with a domain-general executive component and domain-specific storage systems, and overall, these different approaches are complementary in their contributions to SLA (Williams, 2012). Following Williams' (2012) definition, working memory in the current study refers to "the system used for the temporary maintenance of task-relevant information whilst performing cognitive tasks" (p. 456).

With regard to the executive component in working memory, one of the most influential frameworks was proposed by Miyake and colleagues (Miyake & Friedman, 2012; Miyake, Friedman et al., 2000). Their framework focuses on the three most frequently postulated functions of the executive component (Miyake, Friedman et al., 2000): *updating*, which refers to "constant monitoring and rapid addition/deletion of working memory contents", *shifting*, which refers to "switching flexibly between tasks or mental sets", and *inhibition*, which refers to "deliberate overriding of dominant or prepotent responses" (Miyake & Friedman, 2012, p. 9). As Miyake and colleagues pointed out, the three functions were "highly specific and can be defined in a fairly precise manner" (Miyake, Emerson, & Friedman, 2000, p. 177), and can be measured

with “a number of well-studied, relatively simple cognitive tasks” (Miyake, Friedman et al., 2000, p. 55). Adopting a latent variable approach, Miyake and colleagues proposed that these three executive functions have both unity and diversity (Miyake, Friedman et al., 2000), with *unity* referring to the common executive abilities, and *diversity* referring to the updating-specific and the shifting-specific abilities (Miyake & Friedman, 2012).

A recent meta-analysis on working memory capacities and L2 processing and outcomes conducted by Linck et al. (2014) revealed that both the storage and the executive components accounted for individual variances in L2 outcomes, and implied that individual differences in working memory capacities may be associated with both the storage and executive components. Compared with the storage components, especially the phonological short-term memory, the executive component is understudied and more research is needed to specify the relationship between different executive functions and different aspects of L2 learning (Linck et al., 2014).

2.2.2 Operationalization and Measures of Working Memory Capacities

Table 1 summarizes the often-used measures for the functions of working memory. Based on the task complexity, measures used in the working memory literature can generally be divided into simple span tasks and complex span tasks (Colom, Rebollo, Abad, Shih, 2006; Conway et al., 2005; Kane, Hambrick, Tuholski, Wilhelm, Payne, & Engle, 2004; Linck et al., 2014; Miyake et al., 2001; Unsworth & Engle, 2007; Williams, 2012). In simple span tasks, also called short-term memory tasks, participants will first be given a list of items and then be asked to recall all items in the order they are presented (Unsworth & Engle, 2007). A simple span task usually contains several lists, with varying numbers of items in each list (Colom et al., 2006; Unsworth & Engle, 2007; Williams, 2012). Depending on the items to be recalled, simple span

tasks have several variations, including digit span, nonword span, and letter span tasks (Colom et al., 2006; Unsworth & Engle, 2007; Williams, 2012). For example, in a letter span task, participants will first see the letters, *Z*, *J*, *M*, *K*, *X*, and *T*, and then will need to recall these letters in the exact order they are presented.

Table 1. Measures of Working Memory Capacities

Function		Measure
Storage		Digit Span, Nonword Span, Letter Span
Storage & Processing		Reading Span, Operation Span, Counting Span, Symmetry Span
Executive	Updating	Letter Memory, Keep Track, Tone Monitoring, Spatial 2 Back
	Shifting	Number Letter, Plus Minus, Local Global, Color Shape, Category Switch
	Inhibition	Stroop, Anti-Saccade, Stop Signal

Based on the functional importance of concurrent storage and processing in working memory, complex span tasks were developed by inserting a processing task between the items to be recalled in simple span tasks (Colom et al., 2006; Conway et al., 2005; Unsworth & Engle, 2007). For example, in a reading span task, participants will first read a sentence and answer a question about the sentence, and then will see a letter. After a list of sentence-letter pairs are presented, participants will need to recall the letters in their presented orders. A complex span task will also include multiple lists in varying lengths. Complex span tasks originated from Daneman and Carpenter's (1980) reading span task, and were further developed and extended by Engle and colleagues to include operation span, counting span, and symmetry span tasks (Redick, Broadway, Meier, Kuriakose, Unsworth, Kane & Engle, 2012; Unsworth, Heitz, Schrock, & Engle, 2005; see Conway et al., 2005 for methodological review). Given that standard complex span tasks are time-consuming, Engle and colleagues have developed and validated shortened versions of complex span tasks for practical purposes (see Foster, Shipstead, Harrison, Hicks, Redick, & Engle, 2015 and Oswald, McAbee, Redick, & Hambrick, 2014 for

two ways to shorten tasks).

Traditionally, simple span tasks are used to measure the storage function of working memory, while complex span tasks are used to index both storage and processing functions (Conway et al., 2005; Linck et al., 2014; Wlliams, 2012). Specifically, Engle and colleagues' major argument for their Executive-Attention Theory (Engle et al., 1999; Engle & Kane, 2004) was built on complex span tasks as measures of executive control: performing the dual-task of storing and manipulating information will rely mainly on the attentional control ability, and differences in attentional control abilities will lead to differential task performances. Notably, recent re-analyses of datasets in prominent working memory studies (e.g., Engle et al., 1999; Engle & Kane, 2004; Miyake et al., 2001) indicated that simple and complex span tasks may measure processes more similar than previously thought (Colom et al., 2006), and that factors such as scoring methods, list lengths, and presentation modalities may affect the extent to which each process was involved (Unsworth & Engle, 2007). Results of re-analyses showed that simple and complex span tasks had generally comparable correlations with higher-order cognitive abilities (e.g., language comprehension), and it has been suggested that a better strategy is to use multiple tasks to measure working memory capacities (Colom et al., 2006; Unsworth & Engle, 2007), because composite scores of multiple tasks will be less affected by task-specific features and will provide better indices of the shared cognitive processes (Conway et al., 2005; Forster et al., 2015).

Measures of working memory capacities can also be categorized according to the contents (i.e., verbal, visuo-spatial) and languages (first-language [L1], L2). Verbal tasks include digit span, nonword span (see Gathercole, 2006 for review), operation span, and reading span, whereas visuo-spatial tasks consist of matrix span, arrow span, symmetry span and rotation span

(Kane et al., 2004; Miyake et al., 2001). Particularly relevant to SLA research is whether the tasks should be administered in participants' L1 or L2. Several empirical studies have found that L2 working memory scores were correlated with L2 proficiency (e.g., Gass & Lee, 2011; Service, Simola, Metsaenheimo, & Maury, 2002; van den Noort, Bosch, & Hugdahl, 2006), and the recent meta-analysis by Linck et al. (2014) found that when the task language was L2, the correlations between working memory scores and L2 processing and proficiency measures were higher. Linck et al. (2014) explained that the inflated correlations were due to the confounding effects of L2 proficiency, because L2 working memory tasks measured not only working memory capacities but also L2 proficiency. Link et al. (2014) suggested that if the purpose is to isolate working memory capacities from L2 proficiency, the task language should be L1 rather than L2 (see also Gass & Lee, 2011).

In regard to the executive component of working memory, the updating, shifting, and inhibition functions can be measured with relatively simple cognitive tasks respectively (Miyake, Friedman et al., 2000). Tasks for measuring the updating function include letter memory, keep track, tone monitoring, and spatial 2 back tasks. To measure the shifting function, researchers can use number letter, plus minus, local global, color shape, and category switch tasks. For the inhibition function, common measures include Stroop, anti-saccade, and stop signal tasks. (See Miyake, Friedman et al., 2000 and Friedman, Miyake, Young, DeFries, Corley, & Hewitt, 2008 for details about these tasks.) As Miyake, Emerson et al. (2000) recommended, using multiple, simpler tasks can alleviate the idiosyncratic requirements of different tasks and provide a better measure of executive functions.

2.2.3 Working Memory and Attention in L2 Learning

Individual differences in working memory capacities as first proposed by Just and Carpenter (1992) have inspired L2 researchers to speculate working memory as an important component of language aptitude (Williams, 2015). Empirical studies have supported the contributions of working memory to language aptitude (e.g., Li, 2013; Winke, 2013a), and working memory measures have been included in language aptitude tests, such as the High Level Language Aptitude Battery (Hi-LAB, Linck, Hughes, Campbell, Silbert, Tare, Jackson, Smith, Bunting, & Doughty, 2013). It has been found that L2 vocabulary acquisition is correlated with working memory capacities as measured by simple span (e.g., Service & Craik, 1993; Service & Kohonen, 1995; Speciale, Ellis, & Bywater, 2004; Williams & Lovatt, 2003) and complex span tasks (e.g., Kim, Christianson, & Packard, 2015; Martin & Ellis, 2012). Additionally, working memory capacities have been found to be related to L2 grammar learning (e.g., Harrington & Sawyer, 1992; Martin & Ellis, 2012; Robinson, 2002, 2005), effectiveness of recasts (e.g., Révész, 2012; Sagarra, 2007), L2 skill development (e.g., Kormos & Sáfár, 2008), and study abroad experience (e.g., Sunderman & Kroll, 2009), among others (see Williams, 2012 for review; also see Linck et al., 2014 for meta-analysis of working memory and L2 processing and outcomes).

As Williams (2015) pointed out, SLA research should build on current studies that try to discover which aspects of L2 acquisition correlate with which working memory measures, and should move towards using working memory to understand the underlying cognitive processes and to inform SLA theory development. Given the essentiality of executive attention across different approaches to working memory (Baddeley, 2012; Cowan, 2017; Kane, Conway, Hambrick, & Engle, 2007), working memory is important to L2 acquisition because it may affect

cognitive processes during L2 learning by manipulating attentional control (Robinson et al., 2012; Schmidt, 2012; Williams, 2012, 2015). Although the theoretical importance of working memory for attentional control is well discussed and established, empirical studies that directly examined the relationship between working memory and attentional control have produced mixed results.

Some studies have found effects of working memory. Mackey et al. (2002) may be the first to directly address the relationship among working memory, noticing, and L2 development. Working memory capacities were operationalized as the composite scores of L1 (Japanese) and L2 (English) listening span and nonword span tasks, and noticing was operationalized as either verbal reports in stimulated recall or answers to an exit questionnaire. Thirty participants were divided into high, medium, and low capacity groups based on their composite working memory scores. Results of 20 participants in the high and the low capacity groups indicated that participants with high capacities tended to report more noticing of recasts. A subset of data also demonstrated higher capacities were associated with more L2 gains in the delayed posttest, while lower capacities were connected to more L2 development in the immediate posttest. Notably, detailed examination of participants' developmental stages of question formation revealed that the relationship between working memory and noticing was influenced by developmental stages. Among participants with high capacities, those at lower stages were more likely to notice the recasts than those at higher stages.

Another study by Lai, Fei, and Roots (2008) on recasts in a computer-mediated-communication context also found the effects of working memory on noticing of recasts. Working memory capacities were measured by a reverse digit span task, and noticing was measured by think-aloud protocols and stimulated recall. Results of 17 participants showed that

working memory scores had a higher correlation with frequency of noticing non-contingent recasts than with frequency of noticing contingent recasts, while L2 proficiency measured by institute placement tests did not correlate with the frequency of noticing recasts significantly. Kim, Payant, and Pearson (2015) have also found working memory as a significant predictor of noticing of recasts when learning English question formation. Different from traditional measures, an immediate cued recall was used to indicate noticing of recasts. Noticing was operationalized as learners' responses to the immediate cued recall: if learners repeated the recast utterances including the target structure, it was coded as full repetition; if the repeated recast utterances did not include the target structure, it was regarded as partial repetition; if no repetition, it was coded as no repetition. As for working memory capacities, an aural running span task was used. A multiple regression analysis revealed that working memory scores explained 21% unique variance in noticing of recast. Results of three oral production tasks also showed that working memory scores successfully predicted question development.

Other studies did not detect the effects of working memory. Bell (2009) found working memory did not affect awareness of grammatical structures in a crossword puzzle, after controlling learner proficiency. Forty-six participants' working memory capacities were measured with Daneman and Carpenter's (1980) original reading span task, which asked for recall of the final word in the sentence. Awareness was measured with a think-aloud protocol and two probe questions, and was operationalized at two levels, i.e., aware and no verbal reports. A regression analysis did not reveal working memory scores as a significant predictor of awareness. Similarly, in Chen's (2013) study with 60 learners, working memory capacities did not predict noticing of recasts. Working memory capacities were measured by an L1 Chinese reading span task adapted from Daneman and Carpenter's (1980) original reading span task, with

the last two-character word in the sentence to be recalled. Noticing of recasts was measured with a stimulated recall and was operationalized at three levels: noticing content of recasts or noticing other information irrelevant to the corrective nature of recasts, noticing only the corrective function of recasts, and noticing the gap between errors and recasts. Results of simple linear regressions showed that working memory scores did not explain more than 4% of variance in noticing of morphosyntactic or lexical/phonological recasts.

Notably, the aforementioned studies mostly used think-aloud protocols and stimulated recalls to measure noticing or awareness. One recent study by Indrarathne and Kormos (2017) applied the latest eye-tracking technology to measure attention and investigated the relationship between working memory, attention, and L2 grammatical development. Working memory capacities were measured with a digit span task for phonological short-term memory, and a plus-minus, a keep-track, and a Stroop task for the shifting, updating, and inhibition functions of central executive, respectively. Attention was operationalized as total fixation duration (TFD) and the difference between observed and expected TFD in different experimental conditions. Results showed that working memory scores were significantly correlated with attention measures and with learning gains as measured by a sentence reconstruction and a grammaticality judgment task.

Although attention, noticing, and awareness were often used as a post-hoc explanation or as a theoretical premise (Leow, 1999a, 1999b; Truscott, 1998), the aforementioned studies signify progress in that the researchers moved towards empirically examining whether attention, noticing, and awareness affected L2 learning processes and outcomes, and whether their effects were influenced by working memory capacities. Particularly, Indrarathne and Kormos (2017) were innovative in adopting a simple span and three simple executive tasks to measure both the

storage and executive components of working memory, and using eye-tracking to examine the relationship between working memory capacities and online processing and learning outcomes of L2 grammar. Their study responded to Linck et al.'s (2014) call for more research on specifying the connections between executive functions and aspects of L2 learning.

2.3 Presentation Formats and Split-attention

Even before individual differences captured research attention in SLA, L2 input and its external features had received major interest from L2 researchers (Leow et al., 2014). Regardless of theoretical standpoints (e.g., behaviorist, generativist, connectionist, interactionist, and socioculturalist), it is indisputable that input is an indispensable element in SLA; the dispute only lies in its role and extent of importance (Gass, 2010; Gass & Mackey, 2015). Apart from its key role in SLA theories, L2 input also assumes a high status in L2 instruction, especially how to present L2 input in an optimal way to maximize learning (Benati, 2016). As a long-standing topic in instructed SLA, input manipulation has also topped the research agenda of instructional design in educational psychology (Sweller, Ayres, & Kalyuga, 2011). In the following section, I will first overview input manipulation and attention in SLA, then move to educational psychology and introduce the split-attention effect in instructional design, and finally discuss empirical studies on presentation formats and split-attention effects in L2 learning.

2.3.1 Input Manipulation and Attention in SLA

In instructional settings, one major goal is to provide input that facilitates L2 acquisition (Benati, 2016; Lee & Huang, 2008). As mentioned previously, the assumption that attention to input will result in acquisition provides a major rationale for input manipulation, with a focus on

how to improve learner attention to the language target so that better learning outcomes will be attained (Lee & Huang, 2008; Loewen, 2020; Polio, 2007). In SLA research, input manipulation is usually aimed at increasing the salience of target vocabulary and grammatical structures, which often slip from learner attention in meaning-focused contexts (Han, Park, & Combs, 2008). With communicative language teaching gaining popularity among L2 practitioners, focus on form has been proposed to briefly direct learner attention to the linguistic form during meaning-based communication (Long, 1991, 1996). In terms of input manipulation, types of focus on form include input flood and input enhancement (Benati, 2016; Loewen & Inceoglu, 2016). Input flood provides multiple instances of target structures in a meaning-focused context, in the hope that salience in frequency of the target structures will draw learner attention (Hernández, 2011). Input enhancement aims to make the features of the target structures more salient, so that learner attention is attracted to the linguistic form in meaning pervasive contexts (Sharwood Smith, 1981, 1991).

So far, input manipulation in SLA has mainly focused on increasing the salience of the target structures by increasing examplars of the target structure and/or highlighting them in some way (Benati, 2016). Another important component of input manipulation can be the presentation formats of input (Lee & Kalyuga, 2011), which is the focus of instructional design research in educational psychology. SLA research on input manipulation can draw on instructional design research to present L2 input in ways that are in accordance with general principles of effective instructional design. In the upcoming section, I will move to theories of instructional design in educational psychology and discuss their implications for L2 input manipulation.

2.3.2 Cognitive Load Theory and Split-attention Effects in Educational Psychology

One prominent theory of instructional design in educational psychology is the Cognitive Load Theory (Paas, Renkl, & Sweller, 2003, 2004; Sweller, van Merrienboer, & Paas, 1998; van Merrienboer & Sweller, 2005; for latest overview see Sweller, Ayres, & Kalyuga, 2011). According to the Cognitive Load Theory, the human cognitive architecture consists of two major systems: limited working memory and unlimited long-term memory (Sweller et al., 1998; van Merrienboer & Sweller, 2005). Within this framework, human knowledge is stored as schemas in long-term memory, and novel information must be processed by working memory to allow for subsequent schema construction and automation (Paas et al., 2004; Sweller et al., 1998; van Merrienboer & Sweller, 2005). Given the limited capacity of working memory (e.g., Baddeley, 2012) and its key role as a gatekeeper, the major goal of the Cognitive Load Theory is to optimize instructional design so that working memory will not be overloaded to hamper efficient processing (Sweller et al., 1998; van Merrienboer & Sweller, 2005).

Cognitive load refers to the mental effort required of the cognitive system when performing a task, and has three main categories: *intrinsic* load, which results from the levels of element interactivity (i.e., complexity) in the learning materials in relation to learner prior knowledge; *germane* load, which is directly used for schema construction and automation; and *extraneous* load, which is irrelevant to schema construction and automation (Sweller et al., 1998). For learning to take place, the overall cognitive load should not exceed the working memory capacity (Paas et al., 2003). Intrinsic load cannot be altered by instruction directly (Sweller et al., 1998; van Merrienboer & Sweller, 2005); learning to perform *new*, complex cognitive tasks entails high-level interactivity of elements in the learning materials, but as learner expertise develops, intrinsic load will go down with schema construction and automation (Paas et

al., 2003). Conversely, germane load and extraneous load can be manipulated by instruction; instructional techniques for reducing extraneous load have long been the major focus of the Cognitive Load Theory, and research on increasing germane load has been gaining momentum since the last decade (van Merrienboer & Sweller, 2005). To maximize learning, with the inherent intrinsic load depending on learner expertise, the extraneous load should be eliminated as much as possible, in the hope that more germane load will be available and devoted to learning (Paas et al., 2004).

The split-attention effect is well established to provide guidelines for reducing extraneous load in instructional design (Ginns, 2006; Sweller et al., 1998, 2011). When two or more sources of information, essential for understanding but unintelligible in isolation, are presented in a separated format, *split-attention* occurs; learners need to split their attentional resources during learning, with some attentional resources used for mental integration of the disparate sources of information, and other attentional resources used for information processing and schema construction and automation (Sweller et al., 2011). According to the Cognitive Load Theory, mentally integrating information in a split-source format will cause extraneous load; the cognitive resources are not directly used for schema construction or automation (Ginns, 2006; Sweller et al., 1998, 2011). When the materials causing split-attention are converted into an integrated format to eliminate extraneous load and result in *better* learning outcomes, the split-attention *effect* occurs (Ginns, 2006; Sweller et al., 1998, 2011).

Split-attention can be caused spatially or temporally (Ginns, 2006; Sweller et al., 2011). Whereas *spatial* split-attention *effects* (also spatial contiguity effects, Mayer, 2001) are concerned with improved physical layouts of visual information that leads to *better* learning outcomes, *temporal* split-attention *effects* (also temporal contiguity effects, Mayer, 2001) occur

when different sources of information in asynchronous presentations are converted to concurrent presentations that result in *better* learning outcomes (Ginns, 2006; Sweller et al., 2011). Split-source information causing split-attention can come in the form of text and text, text and diagram, or diagram and diagram, with the text in written or spoken modality (Sweller, 1999). Apart from physically integrating split-source information, split-attention can also be eliminated by presenting one of the multiple sources of visual information in audio form, and the *modality* effect occurs when such audio and visual presentations lead to better learning outcomes (Mayer, 2001; Sweller et al., 1998, 2011). It is hypothesized that information presented in different modalities is processed by different components in working memory (e.g., phonological loop and visuo-spatial sketchpad, Baddeley, 2012; see also Wickens, 2008), and working memory capacities can be "expanded" in the sense that more information can be processed in two modalities than in one modality (Mayer, 2001; Sweller et al., 1998, 2011). Another effect closely related to the split-attention effect is the *expertise reversal* effect (Kalyuga, Ayres, Chandler, & Sweller, 2003; Kalyuga, Chandler, & Sweller, 1998; Sweller et al., 2011). The expertise reversal effect highlights the importance of learner prior knowledge: turning split-attention formats into integrated formats may work only for novice learners, but for high-level learners, it may have neutral or even negative effects on learning (Kalyuga et al., 1998, 2003; Sweller et al., 2011). The supportive information for novice learners may become redundant for advanced learners with more prior knowledge, and such redundancy may initiate additional cognitive load that hinders learning (Kalyuga & Renkl, 2010).

A meta-analysis of 50 studies with 2,375 novice students found that split-attention effects are solid and robust, regardless of type of effects (e.g., spatial or temporal), educational level (e.g., primary schools, high schools, or universities), broad field of study (e.g., mathematics,

science, or engineering), or type of information presentation (e.g., static or dynamic) (Ginns, 2006). It also revealed that split-attention effects have a significantly larger effect size for materials with high-level element interactivity (i.e., complexity) than those with low-level complexity, in individual testing than group testing (Ginns, 2006). These results implied that learning materials should be presented in an optimal way to eliminate split-attention for students with little prior knowledge (Ginns, 2006; Sweller et al., 1998, 2011). Particularly, spatial contiguity was well recognized as one of the top 25 "learning principles to guide pedagogy and the design of learning environments" (Graesser, Halpern, & Hakel, 2008) and of the seven recommendations for "organizing instruction and study to improve student learning" (Pashler, Bain, Bottage, Graesser, Koedinger, McDaniel, & Metcalfe, 2007).

Recently, researchers in educational psychology have advocated using eye-tracking to expand the empirical investigation of instructional design, particularly the split-attention effect (e.g., Mayer, 2010; Scheiter & van Gog, 2009; van Gog & Jarodzka, 2013; van Gog, Kester, Nievelstein, Giesbers, & Paas, 2009; van Gog & Scheiter, 2010). Several studies have employed eye-tracking measures such as fixation durations, fixation counts, and transitions to study the split-attention effect in multi-media learning (e.g., Holsanova, Holmberg, & Holmqvist, 2009; Johnson & Mayer, 2012; Mason, Pluchino, Tornatora, & Ariasi, 2013). While *fixation durations* and *fixation counts* are common indices of temporal viewing behaviors and reveal how long the fixations continue and how many different fixations happen, *transitions*, which refer to how many times the eyes leave one area of interest (AOI) and enter another AOI (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka, & van de Weijer, 2011), provide spatial indices about the interaction between instructional design and attention allocation.

The first study that used eye-tracking to examine split-attention is Holsanova et al.'s

(2009) investigation on natural L1 reading of newspapers with text and graphic illustration. In the separated format, the main text and the graphic illustration were placed far from each other, while in the integrated format, the graphic illustration was placed near the relevant content in the main text. Participants' reading behaviors were indexed by *integrative saccades*, which referred to the transitions between semantically related pieces of text and graphic illustration. Results showed that participants made significantly more integrative saccades in the integrated format, which implied that the integrated format facilitated participants' construction of referential connections between two different sources of information. Interestingly, Holsanova et al. (2009) did not find significant correlations between the number of integrative saccades and reading comprehension, and they cautioned that although they interpreted their results as supporting the split-attention effect, another possible interpretation was that the frequent transitions may indicate difficulty in integrating different sources of information. Notably, Holsanova et al. (2009) did not provide any details about how they measured reading comprehension, so it was hard to speculate the real causes of their nonsignificant correlations.

Continuing Holsanova et al.'s (2009) line of research on eye movements and presentation formats, Johnson and Mayer (2012) investigated how different formats of presenting text and diagram affected learning about car brakes. Different from Holsanova et al.'s (2009) definition, *integrative transitions* in Johnson and Mayer's (2012) study referred to the transitions between *any* text and diagram and indicated learners' attempts to integrate different *sources* of information. Additionally, *corresponding transitions* indexed the transitions from the text to the *corresponding* part of the diagram and reflected the extent of successful integration made by learners (integrative transitions in Holsanova et al., 2009). Fixation indices, including fixation counts and total fixation duration, were used to reflect learners' selective attention to the text or

diagram. Learning outcomes were measured with a retention and a transfer test of subjective questions about the car brake system. Results from three sets of experiments showed that participants made significantly more integrative and corresponding transitions in the integrated than in the separated format, and that in both formats, the diagram did not receive more fixations than the text. As for learning outcomes, participants' performance in the transfer test was significantly better with the integrated than the separated format (except for Experiment 3). Johnson and Mayer (2012) interpreted their findings as that the integrated format facilitated meaningful learning by encouraging learner attempts to integrate (i.e., frequent transitions between) different sources of information and by improving integration success (i.e., more corresponding transitions). Additionally, Johnson and Mayer (2012) pointed out that the integrated format did not bias attention to the diagram and that learning was mostly text-driven (i.e., more fixations on the text). Holsanova et al.'s (2009) and Johnson and Mayer's (2012) studies demonstrated that eye-tracking can further reveal the split-attention effect, including why and how it happens (Scheiter & van Gog, 2009; van Gog & Jarodzka, 2013; van Gog et al., 2009; van Gog & Scheiter, 2010).

2.3.3 Presentation Formats and Split-attention Effects in L2 Learning

Aiming to provide general principles for instructional design and facilitate learning across different fields of study, the Cognitive Load Theory, particularly the split-attention effect, has important implications for SLA. Similar to SLA theories, the Cognitive Load Theory highlights the role of working memory in allocating attentional resources during learning, and champions the importance of manipulating input to direct learner attention for efficient learning. The split-attention effect indicates that L2 input manipulation may also need to take into

consideration the effects of different presentation formats, in addition to increasing the salience of the target structures. By considering both linguistic contents and presentation formats of L2 input, L2 research can generate a more comprehensive picture of the relationship between L2 input and acquisition, and can provide inspiration and implications for L2 instruction to facilitate L2 development.

Several studies have investigated split-attention and found effects of presentation formats on L2 learning. The first study to investigate split-attention effects and L2 learning may be Yeung, Jin, and Sweller's (1997) study on L2 English reading comprehension and vocabulary learning with 8th Graders. In their study, the integrated format was created by placing the definitions near the vocabulary in the reading text, while in the separated format, the vocabulary and definitions were placed after the passage. Results showed that with the integrated format, students of lower proficiency gained higher scores for reading comprehension but lower scores for vocabulary knowledge (Experiment 4), whereas more advanced students performed worse in the comprehension test but better in the vocabulary test (Experiment 5). In another study, Yeung (1999) worked with 5th and 8th Graders learning L2 English, and compared the effects of integrated (i.e., inserting vocabulary definitions into the text) and separated (i.e., placing vocabulary definitions at the end) formats on vocabulary learning and reading comprehension. Notably, while Yeung et al. (1997) used the same texts for both beginning and advanced groups, Yeung (1999) used different texts for 5th and 8th Grade learners. Results of Yeung's (1999) study showed a similar pattern to those in Yeung et al. (1997): in the integrated format, 5th Graders had better performance in comprehension but worse performance in vocabulary knowledge (Experiment 1), while 8th Graders showed the reverse performance outcomes (Experiment 2). Both Yeung et al. (1997) and Yeung (1999) suggested that with regard to

presenting learning materials, the split-attention effects should be considered along with task nature (e.g., reading comprehension vs. vocabulary learning) and learner proficiency (e.g., high vs. low). Another study by Marefat, Rezaee, and Naserieh (2016) compared the effects of in-text and marginal glosses, namely, the integrated and separated formats, respectively, on pre-intermediate learners' online reading comprehension in L2 English. Different from previous studies, the vocabulary definitions were not static in Marefat et al.'s (2016) study: they were initially invisible and appeared only after the learner clicked on the highlighted target word in the text. In the in-text condition, the L1 glosses popped up near the L2 vocabulary, while in the marginal condition, the L1 glosses appeared at the right margin. Results of reading comprehension as measured by a multiple-choice test and written recall showed that the in-text glosses (i.e., integrated format) led to statistically significant higher scores.

Apart from changing the locations of vocabulary definitions, the effects of presentation formats on reading comprehension can also be examined by changing the locations of comprehension questions. Following this operationalization, Hung (2007) claimed to have found the split-attention effect on L2 English reading comprehension. In the learning phase, a passage and ten questions to aid comprehension were physically integrated by presenting the questions between the paragraphs in the integrated format, while in the separated format, the questions were placed after the passage. Then in the testing phase, students completed another set of test questions on reading comprehension. Results showed that higher scores were gained with the integrated format in both learning and testing phases. Similarly, Al-shehri and Gitsaki (2010) operationalized the integrated format as presenting questions within the text, and the separated format as placing the questions after the text. They also included another variable, namely, availability of online dictionaries (yes vs. no) in addition to the presentation format variable to

investigate online reading comprehension and vocabulary learning in L2 English. Results showed that availability of online dictionaries had a stronger effect on the comprehension and vocabulary scores than the presentation formats did, but it was also found that less time was needed for reading in the integrated format, which Al-shehri and Gitsaki (2010) interpreted as less cognitive load induced by the integrated format. Adopting the integrated format that places questions within the text and the separated format that presents questions after the text, Genc and Gülözer (2013) investigated the effects of presentation formats and type of presentation (paper-based vs. online) on *advanced* learners' L2 English reading comprehension. They did not find a statistically significant difference in comprehension scores between the two formats, but found statistically higher scores in online reading compared with paper-based reading. These results were in accordance with previous findings that the split-attention effect may be more applicable with novices than advanced learners.

Well established in instructional design research and educational psychology, the split-attention effect is also common in L2 learning (see also Chung, 2007; Lee & Kalyuga, 2011, and detailed discussions of these two studies are in the following section 2.4). If input manipulation is to present L2 input in an optimal way so that attentional resources can be used efficiently for better learning outcomes, then current SLA research can benefit from research on the split-attention effect and the Cognitive Load Theory. By taking both presentation formats and linguistic contents of the input into consideration, L2 research can provide better guidance for effective L2 instruction to facilitate L2 development.

2.4 Working Memory, Presentation Formats, and Learning L2 Chinese Characters

Based on the writing-language relationships depicted by a writing system (Perfetti & Liu,

2005), Chinese is often categorized as a logographic language, or more accurately a morphosyllabic language (DeFrancis, 1989; Mattingly, 1992; Perfetti & Zhang, 1995). Formed by interwoven strokes, a Chinese character functions as a basic orthographic unit and is often mapped to one syllable and one morpheme (Guan, Liu, Chan, Ye, & Perfetti, 2011; Liu, Wang, & Perfetti, 2007; Perfetti, Liu, & Tan, 2005; Wang, Perffeti, & Liu, 2003; Xu, Chang, Zhang, & Perfetti, 2013). A Chinese character can stand alone or be further combined with other characters to form multiple-character words (Perfetti & Liu, 2005; Shen, 2013), e.g., 学 (learn), 学生 (student), 语言学 (linguistics), and 学以致用 (learn for practical purposes). Whereas alphabetic writing systems such as English have generally systematic correspondence between a grapheme (e.g., a letter) and a phoneme, such sub-syllabic correspondence does not apply to Chinese (Liu et al., 2007; Wang et al., 2003). Homophones are also far more common in Chinese than in English (Perfetti et al., 2005; Wang et al., 2003): on average about five Chinese characters share exactly the same pronunciation, with tones taken into consideration (Perfetti et al., 2005). For example, 树 (tree), 述 (tell), 束 (bunch), 漱 (wash), and 竖(vertical) share exactly the same pronunciation. Pinyin was first proposed as a standard phonetic spelling system to facilitate character learning in China (Zhou, 1986), and now has been widely adopted in Chinese language education worldwide (Everson, 2011). The pinyin system uses English alphabetic letters to spell syllables, but pinyin letters have different pronunciation than English letters do (Shen, 2013). Additionally, four distinct tones are differentiated with tonal markers, e.g., /bā/ (high-level), /bá/ (rising), /bǎ/ (low-falling-rising), and /bà/ (high-falling), in addition to a neutral tone, e.g., /bɑ/ (mid-flat) (Liu, Wang, Perfetti, Brubaker, Wu, & MacWhinney, 2011). Generally, learning L2 Chinese characters involves three elements: the shape, the sound, and the meaning (Perfetti et al., 2005; Perfetti & Liu, 2005; Shen, 2013).

Compared with other languages, Chinese has been found to take much longer learning time to achieve the same L2 proficiency level (Jackson & Malone, 2010), and the difficulty may result from cultural and linguistic differences (Everson, 2011). For native speakers of English, Chinese characters remain the greatest challenge in learning Chinese (Hu, 2010; Ke, Wen, & Kotenbeutel, 2001). Research on L2 Chinese characters has found that learning outcomes are related to factors including stroke density (e.g., Ke, 1996), structural configuration and complexity (e.g., Shen & Ke, 2007), radical knowledge (e.g., Xu, Perfetti, & Chang, 2014), learning strategies (e.g., Shen, 2005), instructional methods (e.g., Guan et al., 2011), and Chinese proficiency (e.g., Zhang & Li, 2016).

Recent studies have investigated the relationship between working memory capacities and learning L2 Chinese characters. Kim, Christianson et al. (2015) examined the effects of working memory and L2 proficiency with 70 Chinese nonnative speakers enrolled in beginning and intermediate Chinese courses at the college level. Learning targets were 18 ancient or extremely uncommon Chinese characters in simple structure (without phonetic components). These characters were categorized into visually similar, normal, and distinct pairs, and within distinct pairs, one stroke of each character was artificially bolded. Character knowledge and L2 proficiency were measured by an oral character naming task and a test of previously learned characters, respectively. Spatial and verbal working memory capacities were measured by a rotation span and a reading span task. Notably, the reading span task was conducted in Korean for Korean native speakers, and in English for speakers of English and other languages. Results showed that neither visual distinctiveness nor L2 proficiency affected naming performance, but overall working memory capacities were associated with naming performance. Specifically, higher spatial working memory capacities were connected with better naming performance of

visually distinct characters, whereas verbal working memory capacities were associated with naming of regular characters.

Another study by Kim, Packard, Christianson, Anderson and Shin (2016) reported the results of the same group of participants in Kim, Christianson et al.'s (2015) study learning another set of 18 low-frequency Chinese characters. The target characters were compound characters with a phonetic and a semantic radical, and were categorized into phonetic consistent, semi-consistent, and inconsistent groups based on phonetic consistency. In Kim et al.'s (2016) study, the learning phase as well as the spatial and verbal working memory tests were the same as in Kim, Christianson et al.'s (2015) study, but the testing phase was different: participants named not only trained characters but also untrained real/pseudo-characters. Results showed that only phonetic consistency was significantly associated with naming performance, whereas L1 background or working memory capacities was not, different from Kim, Christianson et al.'s (2015) findings. Kim et al. (2016) attributed their different results to the differences in characters: when learning characters with a phonetic component, the facilitative effects by phonetic cues may override those by working memory observed in learning simple-structure characters.

Other studies have investigated the effects of presentation formats on learning L2 Chinese characters. Chung (2007) studied the split-attention effect by manipulating the adjacency between the character form, the pinyin, and the English translation. In his study (Experiment 1), the three elements were presented simultaneously from left to right on a cardboard in four conditions differing in the order of the elements: character-pinyin-English, character-English-pinyin, English-pinyin-character, and pinyin-English-character. Thirty-two English speaking students in middle school learned 16 two-character Chinese words, with four in

each condition. Vocabulary knowledge was measured by asking students to provide the pronunciation and the meaning of the characters. Two major results were found from an immediate and a delayed posttest: a) when characters were placed first on the left, learning outcomes of meaning and pronunciation were better than when they were presented after pinyin or English; b) when pinyin or English was placed adjacent to characters, its learning outcomes (pronunciation or meaning) were better than when it was separated by the third element (i.e.., English or pinyin). For example, compared with character-English-pinyin, character-pinyin-English was associated with better learning outcomes for pronunciation (because the pinyin is close to the character), but worse for meaning (because the English is separated by pinyin from the character). Chung (2007) explained the second finding with reference to split-attention: when the element (i.e., pinyin or English translation) was separated and far from the characters, learners would need to hold the character information in working memory and then search and match it with the distant element (i.e., pinyin or English translation). The extra processing of holding the information in working memory would cause extraneous load and hinder efficient learning. Results from Chung's (2007) second experiment further supported his speculation of the split-attention effect on learning L2 Chinese characters.

Lee and Kalyuga (2011) continued this line of investigation and compared the effects of two presentation formats: in the *horizontal* format (see Figure 1a), the characters, the pinyin, and the English translation were presented from left to right, while in the *vertical* format (see Figure 1b), the three elements were presented from top to bottom. Seventy-three English native speakers with Chinese heritage were randomly assigned to learn 25 two-character words in either horizontal or vertical format. A multiple-choice test was employed to assess their knowledge of the mappings between character form and meaning, character form and pronunciation, and

meaning and pronunciation. Results showed that participants who learned with the vertical format performed significantly better in the vocabulary test than those learning with the horizontal format. Lee and Kalyuga (2011) explained that the horizontal format caused split-attention whereas the vertical format provided an integrated format for the mutually referring elements, namely, character, pinyin, and English translation. In the vertical format, the corresponding pinyin was placed exactly below each character, so learners would not need to hold the character information in working memory for subsequent search and match with the pinyin, which could reduce or eliminate the extraneous load and facilitate learning (Lee & Kalyuga, 2011). Future investigation was called for the *adjacent* format (see Figure 1c), which was predicted to further reduce or eliminate split-attention and extraneous load, as not only the corresponding pinyin is placed exactly below each character, but also the English translation is just next to the characters, which might spare learners from the search and match between characters and English translation (Lee & Kalyuga, 2011).

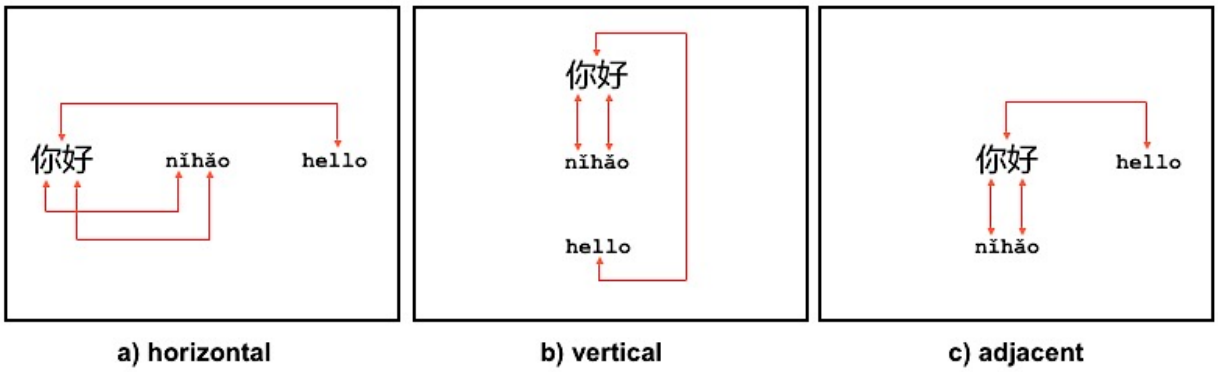

Figure 1. Three presentation formats. Adapted from Lee & Kalyuga (2011).

Notably, both Chung (2007) and Lee and Kalyuga's (2011) used a pre/post-test design to compare the learning outcomes of different presentation formats, and the split-attention effect was used as a post-hoc explanation. Therefore, it remains unclear how learners allocate their attention in different presentation formats *during* the learning process. Additionally, although

both Chung (2007) and Lee and Kalyuga (2011) discussed their findings with reference to working memory, they did not measure learners' working memory capacities, and thus did not provide direct observation of the interaction between working memory capacities and different presentation formats. The current study aims to directly measure attention (via eye-tracking) and working memory capacities, and examine how working memory and presentation formats affect learner attention and learning outcomes of L2 Chinese characters in a computer-assisted self-study environment. Another goal of this study is to expand the line of research on the combined storage and executive components of working memory and L2 grammar learning (e.g., Indrarathne & Kormos, 2017) by exploring these components in L2 vocabulary learning.

2.5 Research Questions and Hypotheses

Following the recent taxonomy of internal and external attention (e.g., Chun et al., 2011), the current study focuses on how two factors, namely, presentation formats (external) and working memory capacities (internal), affect learning L2 Chinese vocabulary in terms of learning outcomes and learner attention in a computer-assisted self-study environment. The overarching research question (RQ) is: What are the relationships among presentation formats, working memory capacities, learner attention, and learning outcomes in L2 Chinese vocabulary learning?

To answer the overarching RQ, two sets of RQs are further proposed to explore the two factors respectively. Specifically, RQ Set A focuses on presentation formats and consists of the following RQs, with each RQ followed by my hypothesis (HP):

1) What is the relationship between *presentation formats* (i.e., horizontal, vertical, and adjacent) and *learning outcomes* (as assessed by a bilingual vocabulary test) in L2

Chinese vocabulary learning, taking L2 Chinese proficiency and vocabulary test formats into consideration?

HP 1: When L2 Chinese proficiency and vocabulary test formats are taken into consideration, the adjacent format will be associated with the highest gain scores, followed by the vertical and the horizontal format.

2) What is the relationship between *presentation formats* (i.e., horizontal, vertical, and adjacent) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in L2 Chinese vocabulary learning?

HP 2: Each presentation format will be associated with a different pattern of data for the attention indices of characters, pinyin, and meaning, i.e., a different combination of large and small numbers for the two attention indices of the three elements in a particular presentation format. Within each presentation format, characters will receive the largest numbers of fixation durations and fixation counts.

3) What is the relationship between *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) and *learning outcomes* (as measured by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency and vocabulary test formats into consideration?

HP 3: When L2 Chinese proficiency and vocabulary test formats are taken into consideration, larger numbers of overall fixation durations and fixation counts will be associated with higher vocabulary scores in the three presentation formats.

4) What is *learners' preference* (as measured by preference ratings) among three *presentation formats* (i.e., horizontal, vertical, and adjacent) in L2 Chinese

vocabulary learning, taking their verbal reports into consideration?

HP 4: When learners' verbal reports are taken into consideration, the adjacent format will have the highest preference ratings, followed by the vertical and the adjacent format.

5) What is the relationship between *learners' preference* (as measured by preference ratings) and *learning outcomes* (as measured by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency, vocabulary test formats, and their verbal reports into consideration?

HP 5: When L2 Chinese proficiency, vocabulary test formats, and learners' verbal reports are taken into consideration, higher preference ratings will be associated with higher vocabulary scores across three formats.

6) What is the relationship between *learners' preference* (as measured by preference ratings) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking their verbal reports into consideration?

HP 6: When learners' verbal reports are taken into consideration, higher preference ratings will be associated with larger numbers of overall fixation durations and fixation counts.

RQ Set B focuses on working memory capacities and consists of the following RQs:

7) What is the relationship between *working memory capacities* (as measured by a storage, a shifting, an updating, and an inhibition tasks) and *learner attention* (to

characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning?

HP 7: Higher working memory capacities will be associated with larger numbers of overall fixation durations and fixation counts.

8) What is the relationship between *working memory capacities* (as measured by a storage, a shifting, an updating, and an inhibition tasks) and *learning outcomes* (as assessed by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency and vocabulary test formats into consideration?

HP 8: When L2 Chinese proficiency and vocabulary test formats are taken into consideration, higher working memory capacities will be associated with higher vocabulary scores.

CHAPTER 3 RESEARCH METHODS

3.1 Overall Design and Operationalization

Table 2 summarizes the overall design of the current study. This study adopted a within-subject design, that is, every participant completed the same experiment procedure. To use a within-subject design was because of practical consideration about participant recruitment and empirical consideration that individual differences exist in eye movements (Henderson & Luke, 2014). Generally, independent variables are presentation formats and working memory capacities, and dependent variables are learning outcomes and learner attention. Notably, learner attention also serves as an independent variable in some data analysis.

Table 2. Overall Design

Research Design	Within-subject
Predictors	Presentation Formats Working Memory Capacities (Learner Attention)
Outcome Variables	Learning Outcomes Learner Attention
Learning Targets	Two-character Chinese Words
Vocabulary Tasks	Recall and Recognition
Working Memory Tasks	Visual Forward Digit Span Letter Memory Number Letter Stroop
Chinese Proficiency Test	HSK Level 1 Reading
Learner Feedback on Presentation Formats	Post-learning Survey Post-learning Interview
Background Knowledge	Background Survey

Following the common practice, working memory is operationalized as a multicomponent system with a domain-general executive and domain-specific storage components (Linck et al., 2014; Williams, 2012). The storage component was measured by a forward digit span task, which is a widely used measure of verbal short-term memory (e.g.,

Bayliss, Jarrold, Gunn, & Baddeley, 2003; Kane et al., 2004; Shahabi, Abad, & Colom, 2014) and was found to be closely related to vocabulary learning (e.g., Atkins & Baddeley, 1998; Baddeley, Gathercole, & Papagno, 1998; Kaushanskaya, Blumenfeld, & Marian, 2011). Specifically, a *visual* forward digit span task was used to avoid the inconsistency in reading rate, intensity, emphasis, and clarity associated with the speaker who reads aloud the digits in the auditory version (Reeves, Schmauder, & Morris, 2000; Silverman, 2007).

As for the executive component, I followed Miyake and colleagues' framework of the three executive functions (Miyake & Friedman, 2012; Miyake, Friedman et al., 2000), namely, updating, shifting, and inhibition. The updating function was measured by a letter-memory task based on Morris and Jones' (1990) experiment, which was adapted in recent research (e.g., Friedman et al., 2008; Miyake, Friedman et al., 2000; Tamnes, Walhovd, Grydeland, Holland, Østby, Dale, & Fjell, 2013). The shifting function was measured by a number-letter task based on Rogers and Monsell's (1995) design, which was adapted in recent studies (e.g., Friedman et al., 2008; Miyake, Friedman et al., 2000; Yow & Li, 2015). The inhibition function was measured by a Stroop task (Stroop,1935), which has stood as a classic test of inhibition in psychology (see MacLeod, 1991 for review) and was adapted in recent studies (e.g., Friedman et al., 2008; Indrarathne & Kormos, 2017; Miyake, Friedman et al., 2000; Tamnes et al., 2013; Yow & Li, 2015). All working memory tasks were performed in participants' L1, namely, English. Task details will be provided in 3.3 Materials.

Given the theoretical and methodological issues associated with the noticing construct, the current study distinguishes between attention and awareness and used eye-tracking as an online quantitative, objective measure of learner attention (e.g., Godfroid et al., 2013; Indrarathne & Kormos, 2016). Two other common eye-tracking measures, namely, fixation

durations (i.e., how long the fixations continue [Holmqvist et al., 2011]) and fixation counts (i.e., how many different fixations happen [Holmqvist et al., 2011]) were used in the current study. Details of the AOIs will be given in 3.4 Procedure.

Although eye-tracking is commended for its quantitative and objective information of attention, it is less informative about the qualitative and subjective aspects of attention (Godfroid et al., 2013; Leow et al., 2014; Scheiter & Van Gog, 2009; Winke, 2013b). Researchers have recommended triangulating eye movement data with subjective verbal reports such as interviews, stimulated recall, and think-aloud protocol to reveal the quality of attention and the cognitive processes involved (Godfroid et al., 2013; Leow et al., 2014; Scheiter & Van Gog, 2009; Robinson et al., 2012; Winke, 2013b). Thus, a survey and an interview were conducted after the learning phase to elicit participants' individual feedback on the learning process, especially their preferences among different presentation formats. Details of the post-learning survey and interview will be provided in 3.3 Materials.

Learning outcomes were measured with a pretest and a posttest on vocabulary knowledge, which belonged to short-term achievement tests (see Nation, 2013). Among many aspects of word knowledge (see Nation, 2013), establishing the form-meaning link is the very first, essential step of developing lexical knowledge (Laufer, Elder, Hill, & Congdon, 2004; Laufer & Goldstein, 2004; Schmitt, 2008). Given my research goal to investigate vocabulary learning at the early stage, the pretest and posttest in this study focused on the mappings between form and meaning (see Figure 2). Additionally, since learning Chinese characters mainly involves studying the shape, the sound, and the meaning (Perfetti et al., 2005; Perfetti & Liu, 2005; Shen, 2013), knowledge of the sound was operationalized as knowledge of the pinyin (e.g., Shen, 2004, 2010; Shen & Ke, 2007), and was examined through the mappings from

characters to pinyin and from meaning to pinyin (see Figure 2). These directions, rather than the reverse directions of mappings (i.e., from pinyin to characters and from pinyin to meaning), were investigated because there are abundant homophones in Chinese (Perfetti et al., 2005; Wang et al., 2003), which makes it difficult to decide whether a non-target response to the pinyin is due to the insufficient knowledge of the target word or the competition from homophones. Another consideration was the quantity of test items, for it would be too many test items if the mappings in the reverse directions were also included (see 3.3 Materials for details of the tests).

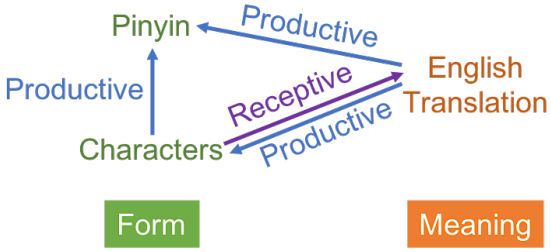

Figure 2. Lexical mappings measured in the pretest and posttest.

Apart from its multiple facets, vocabulary knowledge can also be categorized as receptive or productive: receptive knowledge involves retrieving the meaning of the word form in reading and listening (i.e., meaning recall and meaning recognition), whereas productive knowledge involves producing the word form to express meaning in speaking and writing (i.e., form recall and form recognition) (Nation, 2013). Figure 2 shows the receptive and productive nature of the lexical mappings measured in the pretest and posttest, according to Nation's (2013) categorization of lexical knowledge for assessment. Research has found that receptive knowledge usually develops earlier than productive knowledge (Nation, 2013; Webb, 2008). In other words, going from form to meaning is usually easier than going in the opposite direction. In addition, both receptive and productive knowledge can be measured by recognition (i.e., choose among several options) and recall (i.e., without any options) tasks, with the former

usually being less difficult (Nation, 2013; Webb, 2008). When knowledge (receptive vs. productive) and task (recognition vs. recall) types are considered simultaneously, difficulty level generally climbs from receptive recognition, productive recognition, receptive recall, and finally to productive recall (Nation, 2013).

Focusing on the form-meaning mapping, Laufer and Goldstein (2004) and Laufer et al. (2004) integrated the dichotomies of productive and receptive knowledge as well as recognition and recall tasks and designed four test formats to compare the degrees of strength in lexical knowledge on the form-meaning mapping. The major difference between the two studies was the language of the prompts and options: L1 translation equivalents were used in Laufer and Goldstein's (2004) bilingual version, whereas L2 definitions were employed in Laufer et al.'s (2004) monolingual version. The four test formats in the bilingual version were: 1) form recall (i.e., provide the L2 word form for the L1 translation equivalent); 2) meaning recall (i.e., provide the L1 translation equivalent for the L2 word form); 3) form recognition (i.e., select the L2 word form out of four options for the L1 translation equivalent); and 4) meaning recognition (i.e., select the L1 translation equivalent out of four options for the L2 word form) (Laufer & Goldstein, 2004). Results from both studies generally supported that the difficulty hierarchy of the form-meaning mapping descended from 1) form recall (the most difficult), 2) meaning recall, 3) form recognition, to 4) meaning recognition (the easiest). In the current study, I adapted Laufer and Goldstein's (2004) bilingual version of the four test formats to measure the form-meaning mapping, and also created four parallel test formats (i.e., two recall and two recognition) for the two mappings involving pinyin (see Table 3 for summary). Notably, whether similar patterns in the difficulty hierarchy will be observed in the current study remains an empirical question. Details of the pretest and posttest will be described in 3.3 Materials.

Table 3. Eight Test Formats in Pretest and Posttest

Lexical Mapping	Test Format	Knowledge
From Characters to Meaning	Meaning Recall Meaning Recognition	Receptive
From Meaning to Characters	(Character) Form Recall (Character) Form Recognition	Productive
From Characters to Pinyin	(Pinyin) Form Recall (Pinyin) Form Recognition	
From Meaning to Pinyin	(Pinyin) Form Recall (Pinyin) Form Recognition	

Due to the reality in participant recruitment (see 3.2 Participants for details), a Chinese proficiency test was adapted from the reading component of new HSK Level 1 tests as an additional index to explore potential effects on the results caused by the differences in L2 Chinese proficiency. The new HSK (*Hanyu Shuiping Kaoshi*, Chinese Proficiency Tests) were launched by the Confucius Institute Headquarters affiliated with the Ministry of Education in China in 2009, and are regarded as the most authoritative standardized Chinese exams for non-native speakers (Wang, Zheng, Zheng, Su, & Li, 2016). The new HSK have six proficiency levels for writing and each level was designed to match the Common European Framework of Reference for Languages (CEFR), from A1 to C2 (Hsiao & Broeder, 2013). According to Bachman and Palmer (1996), a language test should correspond to test takers' specific proficiency level in order to maximize test usefulness. Given that participants in the current study had taken college-level Chinese courses for about three, four, or seven months at the time of testing (see 3.2 Participants), I chose HSK Level 1 (equivalent to A1 in CEFR) as the target level, for it was designed for students who mastered 150 words and basic grammar after one semester's classroom instruction (Chinese Testing International, n.d.). Admittedly, for those who had studied Chinese for about seven months, HSK Level 1 may be too simple and HSK Level 2 that aim for two semesters' instruction (Chinese Testing International, n.d.) may be more

appropriate. However, in order to provide the same test index of proficiency, I had to make a compromise and chose HSK Level 1, because it could provide more fine-grained differentiation among those with shorter study times (i.e., three or four months). Another consideration is, if the test is well beyond participants' proficiency levels (i.e., HSK Level 2 for students with three or four months of classroom instruction), it may lead to floor effects and may not distinguish for smaller proficiency differences, such as between three and four months of classroom instruction.

There are two sections in HSK Level 1, namely, listening and reading, with five items for each of the four parts (totally 20 items) in each section (Chinese Testing International, n.d.). Following the practicality principle (Bachman & Palmer, 1996), only the reading section was included in the proficiency test of this study, because it is more relevant to the current focus on the written aspect of vocabulary knowledge. Additionally, the number of test items was decreased to four per part (totally 16 items) in order to reduce the test time and the total time of this study. As mentioned previously, the proficiency test was intended to differentiate among participants with different study times. One way to achieve this was to use stratified sampling of test items according to the syllabus and the textbooks (i.e., *Integrated Chinese Level 1 Volume 1, Lesson 1-10* [Liu, Yao, Bi, Ge, & Shi, 2016], and *Volume 2, Lesson 11-20* [Liu, Yao, Bi, Ge, & Shi, 2017]). By doing this, the proficiency test also served as an achievement test that assessed what students learned in a given syllabus (Davidson & Lynch, 2002). By the time of testing, the three-, four-, and seven- month groups had completed Lesson 5 or 6, Lesson 7, and Lesson 13 in the textbooks, respectively. Therefore, the test items could be grouped into Lesson 1-5, Lesson 6-7, and Lesson 8-13, based on the latest lesson where the vocabulary or grammar that was covered in the test item appeared. To create a pool of test items, I downloaded four sets of official new HSK Level 1 tests online (https://www.digmandarin.com/hsk-practice-test): three sets of past

tests and one set of sample test. After checking the test items in the reading components for the vocabulary and grammar that were covered in the textbooks, I opted for the ratio of including eight, four, and four items from Lesson 1-5, Lesson 6-7, and Lesson 8-13, respectively, based on item availability and participants' study times. Details of the proficiency test will be presented in 3.3 Materials.

3.2 Participants

Originally, 77 students participated in data collection, but 8 were excluded because of their non-English L1s (i.e., Spanish, Malay, Vietnamese) and diagnosed learning difficulties (i.e., dyslexia, attention-deficit/hyperactivity disorder). The 69 participants were English native speakers who did not have Chinese, Korean, or Japanese heritage backgrounds and who had taken elementary-level Chinese courses in college for less than a year in the United States at the time of data collection. Due to the small participant pool and the difficulty of recruiting enough participants, three groups differing slightly in their numbers of months of taking college-level Chinese courses were included (see Table 4). Despite the difference, these participants were regarded as sharing the same beginning level of Chinese proficiency. On average, their self-rated Chinese proficiency was 3.78 for reading, 3.09 for listening, 3.38 for speaking, and 3.04 for writing, out of a 7-point Likert scale (see Table 5 for the descriptive statistics obtained from IBM SPSS Statistics 25). All participants had corrected to normal vision. Except two participants who had color blindness, others did not have color vision deficiency. The two participants will be excluded from data analysis of the Stroop task. Each participant received extra credit (if applicable) and US$35 for completing the study.

Table 4. Participant Groups

	n	Months
Group 1	20	7
Group 2	5	4
Group 3	44	3

Table 5. Descriptive Statistics of Self-rated Proficiency for Four Chinese Language Skills

					95% *CIs*	
	Mean	*SD*	*Min*	*Max*	*Lower*	*Upper*
Reading	3.78	1.40	1	6	3.45	4.12
Listening	3.09	1.17	1	6	2.81	3.37
Speaking	3.38	1.25	1	6	3.08	3.68
Writing	3.04	1.43	1	6	2.70	3.39

3.3 Materials

3.3.1 Target Words

Each participant studied the same 30 two-character Chinese words (see Table 6). The target words were selected from *A Frequency Dictionary of Mandarin Chinese* (Xiao, Rayson, & McEnery, 2009), which includes the 5,004 most commonly used Chinese words based on a 50-million-word corpus of spoken and written texts. Specifically, target word selection was started with the end of the frequency index (i.e., the lowest frequency rank), and all target words were checked to ensure each character was not in the two textbooks of elementary Chinese (i.e., *Integrated Chinese Level 1 Volume 1,* Lesson 1-10 [Liu et al., 2016], and *Volume 2,* Lesson 11-20 [Liu et al., 2017]). All target words consisted of simplified compound characters. They were assigned into three ten-word groups, matched in word frequency, number of strokes, structural configuration, and part of speech (see Appendix A for detailed information of these four indices as well as the pinyin and English translation of each word). Specifically, each word's frequency and part of speech were obtained from Xiao et al. (2009), each word's number of strokes was

collected from an online stroke checking system (https://bihua.911cha.com/), and the structural configuration of each character in the word was recorded from the online *Xinhua Dictionary* (https://zidian.wenku1.com/). Table 7 shows the distribution of part of speech (i.e., noun, verb, and adjective) and structural configuration (i.e., left-right, top-down, and half-enclosure, based on Shen's [2013] categorization) within each word group. Table 8 presents the descriptive statistics of word frequency and number of strokes for each word group (obtained from IBM SPSS Statistics 25). Results of bootstrapped one-way ANOVA and Tukey post hoc tests (from IBM SPSS Statistics 25) found no significant differences among the three word groups in terms of frequency, $F(2, 27) = .045$, $p = .956$, and number of strokes, $F(2, 27) = 1.176$, $p = .324$.

Table 6. Target Words in Three Groups

a	b	c
贫穷	恶劣	姿态
夺取	驾驶	尖锐
忽略	审判	奖励
暗示	欺负	神奇
挑选	强迫	叙述
弥补	指挥	讽刺
形状	细致	伴随
脆弱	消耗	谈论
循环	挖掘	抽烟
流传	珍惜	耽误

Table 7. Distribution of Part of Speech and Structural Configuration Within Each Word Group

Part of Speech			Structural Configuration				
Noun	Adjective	Verb	TD+TD	TD+LR	LR+TD	LR+HE	LR+LR
1	2	7	1	2	1	1	5

Note. There were 10 words in a group. TD = top-down. LR = left-right. HE = half-enclosure.

The three word groups were further checked for the radicals in the characters, as research has found that radicals, including phonetic radicals (i.e., radicals with similar

pronunciations to the whole characters), can provide efficient learning cues (Shen, 2013; Shen & Ke, 2007). To ensure that all three groups of target words also matched in the number of radicals that were covered in the textbooks, I checked the radicals of every target character for whether they were also included in the characters that were covered in the textbooks. If so, I recorded the textbook characters and the lessons where they appeared. Generally, each target word group had about four phonetic radicals and eight non-phonetic radicals covered in the textbooks. (See Appendix B for the details of the target characters and the textbooks characters.) Participants' different study times were also taken into consideration so that each target word group had a similar number of radicals covered in the lessons taught beyond four months of classroom instruction.

Table 8. Descriptive Statistics of Word Frequency and Number of Strokes for Three Word Groups

Group	*n*	Word Frequency			Number of Strokes		
		Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
a	10	73.2	9.319	[58, 88]	16.9	2.378	[14, 20]
b	10	72.0	9.201	[61, 87]	18.1	1.912	[15, 20]
c	10	72.5	8.475	[59, 86]	17.0	1.414	[14, 19]

Note. Word frequency refers to frequency per million words.

Another consideration in matching the word groups was difficulty levels. A Latin square design was adopted (see following paragraphs for details) to counterbalance the presentation of all word groups, which can effectively reduce the confounding effects of nuisance variables including difficulty levels (Loewen & Plonsky, 2015; Tavakoli, 2012). From the perspective of language assessment, it is also helpful to have a general idea of the difficulty levels of test items before the test is administered to the target population (Davidson & Lynch, 2002; Green, 2013). The pilot data I collected (see Appendix C for details) enabled

me to have a rough estimation of the average difficulty levels of each word group, and results showed the word groups had similar difficulty levels. By considering word frequency, number of strokes, structural configuration, part of speech, learned radicals, and difficulty levels, the final version of the word groupings can be regarded as well matched.

With the final version of the word groupings, three word lists were created to counterbalance the presentation formats for all word groups according to a Latin square design (see Table 9). Within each list, the three word groups were in different presentation formats, and across the three lists, each word group rotated among the three presentation formats (see Appendix D for the presentation format of each word in each word list).

Table 9. Word Lists with Different Combinations of Word Groups and Presentation Formats

Group / List	a	b	c
i	Horizontal	Vertical	Adjacent
ii	Vertical	Adjacent	Horizontal
iii	Adjacent	Horizontal	Vertical

3.3.2 Pretest and Posttest

Test Formats. I used Qualtrics (www.qualtrics.com) to create an online pretest and posttest (see Supplementary Materials A for the print version of the complete test). There was no time limit for completing the tests, but on average, participants spent 10 to 15 minutes on the pretest and about 30 minutes on the posttest. This time difference was not surprising, because all target words were carefully chosen to be unknown to most participants, which meant that in the pretest, participants may just need to choose "No" (for recall items, see Table 11 for details) or "I don't know" (for recognition items, see Table 10 for details) without trying hard to come up with the answers.

The pretest consisted of two task categories based on the starting element of the lexical mappings: a) from ***meaning*** to characters and pinyin, and b) from ***characters*** to meaning and pinyin. Under each task category, there were a recall and a recognition tasks for each lexical mapping: i) from ***meaning*** to characters, ii) from ***meaning*** to pinyin, iii) from ***characters*** to meaning, and iv) from ***characters*** to pinyin. Table 10 shows the sample items of the Chinese word 贫穷 (*poor*) in the eight test formats. During the pretest, the test items of the same task category for each target word were presented together as a block, with two recall or two recognition items on the same page. After all test items of the same task category were completed, those of the other task category would appear. In other words, participants would see Items 1 and 2 on the first page, and Items 3 and 4 on the second page. Then they would work on other four test items in these same formats for another word. That is, after Items 1 to 4 of all target words were finished, participants would move to Items 5 to 8 for each target word. Notably, participants were not allowed to go back to a previous page in order to avoid the facilitation of the recognition tasks to the recall tasks. This order of presenting test items was in accordance with Nation's difficulty ranking for productive and receptive knowledge and recall and recognition tasks (2013). In total, there were 240 test items in 60 blocks of two categories for 30 target words in the pretest. The pretest and the posttest were the same, with the blocks of test items randomized within each task category for each participant.

Table 10. Sample Test Items of the Chinese Word 贫穷 That Means Poor

Category (a) From Meaning to Characters and Pinyin		
Lexical Mapping	Task	Knowledge
Item (1) From Meaning to Characters (M2C)	Recall (RCL)	Productive
In this item, participants would see an English translation and would need to handwrite the Chinese characters by using the mouse. (See Figure 3)		

Table 10 (cont'd)

Lexical Mapping	Task	Knowledge
Item (2) From Meaning to Pinyin (M2P)	Recall (RCL)	Productive
In this item, participants would see an English translation and would need to type in the pinyin including tones* for the Chinese word.		
poor: Pinyin (with Tones) Production		
Item (3) From Meaning to Characters (M2C)	Recognition (RCG)	Productive
In this item, participants would see an English translation and would need to choose the Chinese word (see also Lee & Kalyuga, 2011).	poor: Character Recognition I don't know. 夺取 贫穷 讽刺 欺负	
Item (4) From Meaning to Pinyin (M2P)	Recognition (RCG)	Productive
In this item, participants would see an English translation and would need to choose the pinyin with tones for the Chinese word.	poor: Pinyin Recognition I don't know. fěngcì zhēnxī xìzhì pínqióng	
Category (b) From Characters to Meaning and Pinyin		
Item (5) From Characters to Meaning (C2M)	Recall (RCL)	Receptive
In this item, participants would see a Chinese word and would need to type in the English meaning for the word (see also Shen, 2004, 2010; Shen & Ke, 2007).		
贫穷: Meaning Production		
Item (6) From Characters to Pinyin (C2P)	Recall (RCL)	Productive
In this item, participants would see a Chinese word and would need to type in the pinyin including tones* for the word (see also Shen, 2004, 2010; Shen & Ke, 2007).		
贫穷: Pinyin (with Tones) Production		
Item (7) From Characters to Meaning (C2M)	Recognition (RCG)	Receptive

Table 10 (cont'd)

In this item, participants would see a Chinese word and would need to choose the English meaning for the word (see also Lee & Kalyuga, 2011; Shen, 2010).	贫穷: Meaning Recognition I don't know. delay treasure bully poor

Lexical Mapping	Task	Knowledge
Item (8) From Characters to Pinyin (C2P)	Recognition (RCG)	Productive
In this item, participants would see a Chinese word and would need to choose the pinyin with tones for the word.	贫穷: Pinyin Recognition I don't know. dānwù pínqióng xìzhì zhēnxī	

Note. *For tone typing, five numbers were used to represent different tones: 0 (mid-flat), 1 (high-level), 2 (rising), 3 (low-falling-rising), and 4 (high-falling) (Liu et al., 2011).

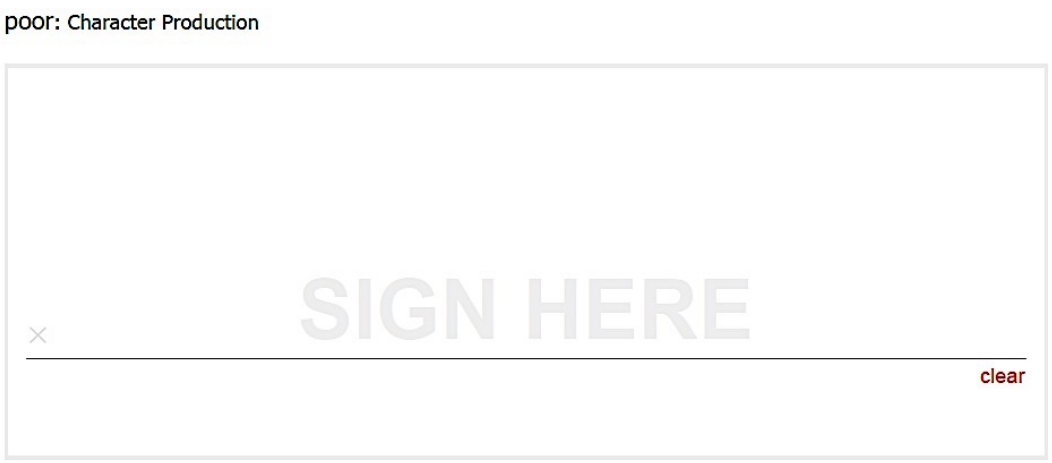

Figure 3. Item (1) from meaning to characters (M2C).

Given the large number of test items and the fact that participants were not very likely to complete the recall items after only short exposures (i.e., totally 34 seconds) to the target words, I added a screening item before the two recall items in each block (see Table 11 for two types of

screening items) by using Qualtrics' adaptive function so as to streamline the test process. Therefore, within the same task category (i.e., starting from *meaning* [Category a] or *characters* [Category b]), participants would first see the screening item for the two recall items for the word. If they chose "Yes" for the screening item, they would then be shown the two recall items on the same page; otherwise, they would not be shown the recall items and move directly to the recognition items. As a result, if participants chose "No" for the screening item, they would automatically receive 0 points for the two corresponding recall items, because they would not see these recall items at all and thus have no chance to work on them.

Table 11. Samples of Two Types of Screening Items Before Recall Items of the Chinese Word 贫穷 That Means Poor

Lexical Mapping	Task	Item Order
Category (a) From *Meaning* to Characters and Pinyin	Recall (RCL)	Before Item 1
poor: Do you know the Chinese characters? No Yes		
Category (b) From *Characters* to Meaning and Pinyin	Recall (RCL)	Before Item 5
贫穷: Do you know this Chinese word? No Yes		

Options for Recognition Items. The recognition items were in the multiple-choice format. For multiple-choice tests, the relationship between the target option and the distractors could be manipulated as overlapping to test precise knowledge (i.e., non-sensitive multiple-choice), or as unrelated to assess imprecise knowledge (i.e., sensitive multiple-choice) (Nation, 2013). Following Laufer et al. (2004), Laufer and Goldstein (2004), and Schmitt, Schmitt, and Clapham (2001), I created sensitive recognition items whose distractors had little overlap with

the target option, so as to measure partial knowledge that resulted from short exposures of vocabulary learning. Specifically, I adopted Lee and Kalyuga's (2011) method of using other target words as distractors, because all target words were not closely related in terms of character form, pinyin, or meaning.

Since the recognition items for a target word were in four formats, namely, (3) From Meaning to Characters, (4) From Meaning to Pinyin, (7) From Characters to Meaning, and (8) From Characters to Pinyin (see Table 10), four versions of the option set, which consisted of three distractors plus one target, were created for each target word with controlled randomization that satisfied the following three conditions. First, each option set for each target word was unique *within* each test format. In other words, the same combination of the four options would appear only once within the same test format. Second, the same combination of two options would appear only twice *within* each test format. That is, every option set would have only two overlapping options with other option sets in the same test format. Third, *across* the four test formats for the target word, no option sets were exactly the same. This meant that for an option set, only some but not all of its options would be shared *between* test formats, given the difficulty in creating option sets that had completely different options for the target word across all test formats. One example to illustrate these conditions is the option set for 贫穷 (*poor*) in Sample Item (7) From Characters to Meaning – *delay*, *treasure*, *bully*, and *poor* (see Table 10). According to Condition 1, these four English words would not appear together as options for other test items in the same test format. Based on Condition 2, other test items in the same test format would include only one of the following two-word combinations as options: *delay-treasure*, *delay-bully*, *delay-poor*, *treasure-bully*, *treasure-poor*, and *bully-poor*. As for Condition 3, the option sets that were defined as the same as the current option set (i.e., *delay*,

treasure, *bully*, and *poor*) were those consisting of the corresponding Chinese characters (i.e., *耽误*, *珍惜*, *欺负*, and *贫穷*) or pinyin (i.e., *dānwù*, *zhēnxī*, *qīfu*, and *pínqióng*) of these English words. With this way of defining the same option sets, Condition 3 required that at least one option was different in the option sets between two test formats for the target word. One instance to fulfill Condition 3 can be: for the target word *贫穷* (*poor*), the option set for Item (7) From Characters to Meaning consists of *delay*, *treasure*, *bully*, and *poor*, and the option set for Item (8) From Characters to Pinyin consists of *dānwù* (*delay*), *zhēnxī* (*treasure*), *xìzhì* (*careful*) and *pínqióng* (*poor*), with *bully* in Item (7) and *xìzhì* (*careful*) in Item (8) as the different options. Apart from the three conditions for the controlled randomization, the order of the options in each option set was randomized for each participant each time. These randomization procedures were aimed to enhance the quality of test items and to reduce the effects unrelated to participants' knowledge of the target words.

For multiple-choice tests, besides the consideration of distractors, another important issue is guessing, which will weaken the argument that participants choose the correct answer because they have the knowledge (Nation, 2013; Schmitt et al., 2001). One way to cope with the guessing issue is to ask participants to leave the test item blank if they have no idea about the answer (Schmitt et al., 2001). This was adapted into the current pretest and posttest by adding an "I don't know" option to reduce participants' intention to guess. However, as Schmitt et al. (2001) acknowledged, it was possible that some participants would still guess and in fact, several participants who attended the interview with Schmitt et al. (2001) did admit that they guessed the answers when working on the test. As the reality of multiple-choice tests, the possibility of guessing could not be completely eliminated from the pretest and posttest of this study.

Test Instructions. Before the pretest, I assured participants that it would be totally fine if

they did not know the answers to the test items, and that their performances would not reflect their Chinese knowledge or proficiency. They were also encouraged to follow the instructions closely and try their best. During the tests, an instruction block would appear at the beginning of each task category, followed by the testing blocks. In the instruction block, participants were first given detailed descriptions of the dos and don'ts of the test with illustrations of sample test items. Then they would try out practice items for one familiar word, which allowed them to check their understanding of the instructions. Notably, participants were reminded that they should make sure they understood what to do with the practice items, because the detailed instructions would not appear with the real test items later. Such design was aimed to save participants' time by avoiding reading the detailed instructions repetitively for each test item. In addition, as mentioned previously, guessing is one major issue with recognition items, so I adapted Schmitt et al.'s (2001) explicit instructions to discourage guessing as follow: "If you have no idea, please do **NOT** guess and choose 'I don't know'. If you think you might know the answer, you should try and find the answer." The detailed instructions for each task category can be found in Supplementary Materials A (pp. 1-2, 23-25).

Item Scoring. Table 12 and Table 13 summarize the different scoring systems for the recognition items and recall items respectively. Specifically, one scoring system was shared by all recognition items, whereas three different scoring systems were employed for the four test formats of recall items. Specifically, one example of the pinyin answers, *jin1qong2*, for the word *贫穷 (poor)* can be used to illustrate the scoring for Items (2) and (6). As the correct pinyin answer is ***pin2qiong2***, $\frac{1}{6}$ points were deducted three times for the incorrect pinyin initial (j vs. p), tone (1 vs. 2), and pinyin final (ong vs. iong), resulting in the loss of 0.5 points.

Test Reliability. As mentioned previously, the pretest and posttest were exactly the same

and the pretest was expected to generate scores approaching 0, so test reliability was calculated based on gain scores. Following the recommendation of calculating test reliability separately for different constructs (Field, 2018), reliability was calculated for each of the eight test formats with IBM SPSS Statistics 25. Specifically, Cronbach's alpha was calculated for the recognition items. As for the recall items, another Chinese native speaker and I graded all items separately and inter-rater reliability, namely, Intraclass Correlation Coefficient (ICC), was calculated. Then we discussed and resolved the grading discrepancy by reaching 100% agreement on the revised grading. Table 14 presents the reliability statistics for the eight test formats. Ranging from .834 to 1.000, the reliability statistics exceeded the generally acceptable value of .70 (Field, 2018) and indicated the test items were of good reliability.

Table 12. Scoring of Recognition Items

Test Format	Score/Item*	Total*
Item (3) From Meaning to Characters (M2C_rcg)	1	30
Item (4) From Meaning to Pinyin (M2P_rcg)	1	30
Item (7) From Characters to Meaning (C2M_rcg)	1	30
Item (8) From Characters to Pinyin (C2P_rcg)	1	30

Note. *The score unit is one point.

Table 13. Scoring of Recall Items

Test Format	Score* /Unit	Unit Total	Item Subtotal*	Score Total*
Item (1) From Meaning to Characters (M2C_rcl)	0.5 /Character[1]	2	1	30
Item (2) From Meaning to Pinyin[2] (M2P_rcl)	$\frac{1}{6}$ /Pinyin Initial	2	1	30
	$\frac{1}{6}$ /Pinyin Final	2		
	$\frac{1}{6}$ /Tone[3]	2		
Item (5) From Characters to Meaning (C2M_rcl)	1 /English Translation	1	1	30
Item (6) From Characters to Pinyin[2] (C2P_rcl)	$\frac{1}{6}$ /Pinyin Initial	2	1	30
	$\frac{1}{6}$ /Pinyin Final	2		
	$\frac{1}{6}$ /Tone[3]	2		

Note. *The score unit is one point.
[1] Only when the whole character was produced correctly would the character be coded as correct; otherwise, it would be coded as incorrect.
[2] For the pinyin syllable of the character 略 (*lüè*) in the word 忽略, either *lue* or *lve* would be marked as correct, because the pinyin letter *ü* does not correspond to any English letter and either *u* or *v* can be used when typing Chinese characters with the pinyin method.
[3] For tone typing, five numbers were used to represent different tones: 0 (mid-flat), 1 (high-level), 2 (rising), 3 (low-falling-rising), and 4 (high-falling) (Liu et al., 2011).

Table 14. Reliability Statistics for Eight Test Formats

Index	Test Item[1]	Reliability Index	Statistics
M2C_rcl	(1) From Meaning to Characters	ICC	1.000*
M2P_rcl	(2) From Meaning to Pinyin	ICC	.989*
M2C_rcg	(3) From Meaning to Characters	Cronbach's alpha	.918
M2P_rcg	(4) From Meaning to Pinyin	Cronbach's alpha	.834
C2M_rcl	(5) From Characters to Meaning	ICC	.998*
C2P_rcl	(6) From Characters to Pinyin	ICC	1.000*
C2M_rcg	(7) From Characters to Meaning	Cronbach's alpha	.938
C2P_rcg	(8) From Characters to Pinyin	Cronbach's alpha	.874

Note. $n = 69$. ICC = Intraclass Correlation Coefficient.
*$p < .05$.
[1] For details of the test items, see Table 10.

3.3.3 Working Memory Tasks

The storage component of working memory was measured with a visual forward digit span task, and the updating, shifting, and inhibition functions of the executive component were measured with a letter-memory, a number-letter, and a Stroop tasks, respectively, in L1 English. All tasks were programmed with E-Prime 2.0 (Psychology Software Tools, 2012) and were completed on a 14' PC laptop with 1920*1080 resolution (Dell Latitude E7470).

Visual Forward Digit Span Task. This task was developed based on the task description in Ostrosky-Solís and Lozano (2006) and Olsthoorn, Andringa and Hulstijn (2014), and followed the guidelines of the *Wechsler Adult Intelligence Scale III Digit Span* subtasks (Wechsler, 1997) (see Figure 4). In this task, participants would see lists of digits, with each digit appearing one by

one on the screen. At the end of each list, participants would need to recall the digits in the order presented. The number of digits in a list (i.e., list length) ranged from 3 to 9, and no digits appeared twice within each list. For each list length, there were two different lists. This task began with two lists of three digits, then moved to two lists of four digits, and so on, with a maximum of 14 lists. However, the task would stop if participants failed to recall two consecutive lists of the same length, and their digit span would be the maximum list length for which they could recall all digits correctly. The digits were presented in 36 Consolas font (bolded, black) at the center of a white background, and each digit stayed on the screen for 1000 millisecond (ms) before the next digit appeared. Specifically, the 1000 ms interval was the same as in Ostrosky-Solís and Lozano (2006) and Olsthoorn et al. (2014). After all digits of a list were presented, participants would have unlimited time to type in the digits with a keyboard. Before starting the real task, participant practiced recalling four lists of two or three digits, with two lists for each length. As mentioned previously, a participant's digit span would fall between 3 and 9. The digit lists used in this task are attached in Appendix E.

Letter Memory Task. This task was developed based on the task description in Friedman et al. (2008), Miyake, Friedman et al. (2000), and Tamnes et al. (2013) (see Figure 4). In this task, participants would see lists of letters (i.e., consonants), with each letter appearing one by one on the screen. For each letter list, participants would need to keep rehearsing and updating the last four letters presented, by mentally adding the most recent letter and dropping the fifth letter back, and speaking out loud the new string of the four letters. At the end of each list, participants would have unlimited time to recall and type in the last four letters in the presented order with a keyboard. For example, if the letter list was J, D, M, K, F, Z, N, participants would need to speak aloud "J… J-D… J-D-M… J-D-M-K… D-M-K-F… M-K-F-

Z… K-F-Z-N", and then type in "KFZN". If participants did not know the letter, they were instructed to say "blank" as a replacement. Within each list, no letters appeared twice. The number of letters in each list (i.e., list length) was 5, 7, 9, or 11, and varied randomly across lists. In this task, participants would first practice four lists, two of which with 5 letters and the other two with 7 letters. Then they would complete 12 lists, with three lists for each list length and a total of 48 letters to recall. The letters were presented in 36 Consolas font (bolded, black) at the center of a white background, and each letter stayed on the screen for 2,500 ms before the next letter appeared. This time limit was the same as Friedman et al. (2008). Participants' oral rehearsal were recorded for confirmation purposes, and their task performances were indexed by the proportion of recalling the letters correctly. Specifically, although participants were instructed to rehearse and recall the letters in the order presented, they would receive points even if the letter order was incorrect, following Friedman et al. (2008). The letter lists used in this task are attached in Appendix F.

Number Letter Task. This task was created according to the task description in Friedman et al. (2008), Miyake, Friedman et al. (2000), and Rogers and Monsell (1995) (see Figure 4). In this task, a large square divided into four sub-squares were presented at the center of the screen, and a number-letter (e.g., 9G) or letter-number pair (e.g., G9) would appear in one of the four sub-squares. When the pair was in one of the top sub-squares (i.e., top-left or top-right), participants would need to indicate whether the number was even (i.e., 2, 4, 6, or 8) or odd (i.e., 3, 5, 7, or 9), by pressing the left key with a left finger (for even numbers) or the right key with a right finger (for odd numbers) on the keyboard. When the pair was in one of the bottom sub-squares (i.e., bottom-left or bottom-right), participants would need to indicate whether the letter was a consonant (i.e., G, K, M, or R) or vowel (i.e., A, E, I, U), by pressing the left key

with a left finger (for consonants) or the right key with a right finger (for vowels) on the keyboard. Specifically, the "C" key was relabeled as the left key "L" and the "M" key as the right key "R" on the US keyboard. There were three blocks of target trials, with each of the first two blocks having 32 trials and the last block having 128 trials. The number-letter or letter-number pair appeared randomly in one of the top sub-squares in the first block, and in one of the bottom sub-squares in the second block. In the third block, the number-letter or letter-number pair appeared in all four sub-squares in a clockwise rotation. To complete the third block of target trials, participants would need to shift between number and letter categorization, but they would not need to do so for the first two blocks. Specifically, mental shifting would be needed when the number-letter or letter-number pair appeared in the top-left or the bottom-right sub-square. The number-letter or letter-number pairs were presented in 36 Consolas font (bolded, black) on a white background. Each pair stayed on the screen until the participants pressed a key, 150 ms after which, another pair would appear. Notably, the time interval was the same as in Miyake, Friedman et al. (2000). Before working on the target trials in each block, participants completed 12 practice trials. Participants' task performances were indexed by the reaction time (RT) differences between the average RTs of the target trials that required mental shifting (i.e., when the number-letter or letter-number pair appeared in the top-left or the bottom-right sub-square in the third block) and the average RTs of the trials in the first and second blocks that did not require mental shifting. The number-letter and letter-number pairs used in this task are attached in Appendix G.

Stroop Task. This task was created according to the task description in Friedman et al. (2008), Miyake, Friedman et al. (2000), and Tamnes et al. (2013) (see Figure 4). In this task, participants would first see a white fixation cross on a black screen, which was then replaced by

a string of asterisks or a color word (i.e., *red*, *blue*, *green*, *orange*, *yellow*, or *purple*). They would need to indicate the ink color (i.e., red, blue, green, orange, yellow, or purple) of the asterisks or the word. Trials of this task were divided into three conditions: 1) control (i.e., the asterisk string was in one of the six ink colors [i.e., red, blue, green, orange, yellow, or purple] and matched the length of one of the six color words [i.e., three, four, five, or six], e.g., ****); 2) congruent (i.e., the color word and the ink color were matched, e.g., *blue*); and 3) incongruent (i.e., the color word and the ink color did not match, e.g., *green*). Participants were instructed to make their responses as accurately and quickly as possible, by pressing the corresponding key on the keyboard (i.e., "R" for red, "B" for blue, "G" for green, "O" for orange, "Y" for yellow, and "P" for purple). Notably, six keys on the US keyboard were relabeled with stickers made of a black letter on a white background: "X" relabeled as "R" (red), "C" relabeled as "B" (blue), "V" relabeled as "G" (green), "B" relabeled as "O" (orange), "N" relabeled as "Y" (yellow), and "M" relabeled as "P" (purple). This task started with 10 practice trials and then 144 target trials (i.e., 72 control, 12 congruent, and 60 incongruent), same as in Miyake, Friedman et al. (2000). Specifically, the target trials were randomized in such a way that neither the ink colors nor the color words (if any) of two consecutive trials was related. For example, the following two-trial sequences were excluded from the task: *green-purple*, *green-purple*, ***-*blue*, and *red-red*. During the task, the fixation cross stayed in the middle of the screen for 500 ms, and then the asterisks or color word appeared and stayed until the participants pressed the keyboard, after which the screen remained black for 1000 ms before another fixation cross appeared. The asterisks and color words were presented in 26 Consolas font (bolded) at the center of the screen. Participants' task performances were indexed by the RT differences between the control trials and incongruent trials. The trials used in this task are attached in Appendix H.

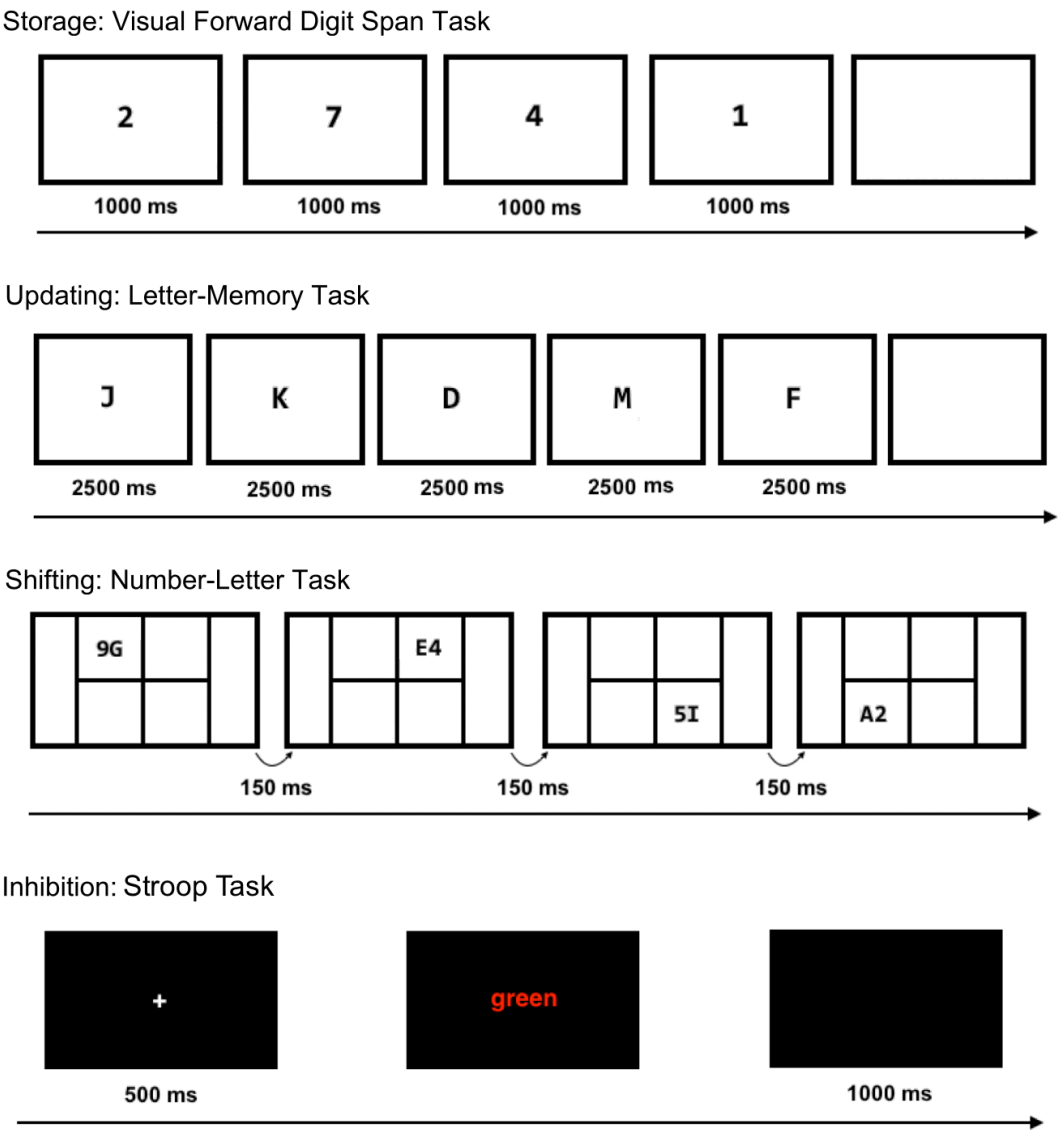

Figure 4. Four working memory tasks.

3.3.4 Chinese Proficiency Test

I used Qualtrics (www.qualtrics.com) to create an online Chinese proficiency test (see Supplementary Materials B for the print version of the complete test). This test was adapted from the reading component of HSK Level 1: the four test formats were the same, but the number of test items for each test format was reduced from 5 to 4 in order to shorten test time. Accordingly, the original time limit of 17 minutes for 20 test items was adjusted to 12 minutes for 16 test

items. To assist time management during the test, a countdown timer was inserted at the beginning, in the middle, and at the end of the test page, respectively. The test items were extracted from four sets of official HSK Level 1 tests downloaded from online resources (https://www.digmandarin.com/hsk-practice-test). Specifically, given participants' different study times (i.e., three, four, or seven months) in taking college-level Chinese courses (see 3.2 Participants), stratified sampling was conducted to select the test items: 8, 4, and 4 test items were chosen for the ranges of Lessons 1-5, Lessons 6-7, and Lessons 8-13, respectively, based on the most advanced vocabulary or grammar covered in the test item (see 3.1 Overall Design and Operationalization). (Also see Appendix I for details of the test items.) As for scoring, one point was awarded to one correct choice and the total score was 16 points. Test reliability was calculated with data from 69 participants with IBM SPSS Statistics 25. Cronbach's alpha was 0.858 and exceeded the generally acceptable value of .70 (Field, 2018), indicating the test items were of good reliability.

3.3.5 Post-learning Survey and Interview

The online post-learning survey was created with Qualtrics (www.qualtrics.com) (see Supplementary Materials C for the print version of the complete survey). There were two questions in the survey: a) one question asked whether the participant noticed something about the presentation formats of the Chinese characters, pinyin, and meaning during learning, and b) the other asked for his or her preference among the three presentation formats (i.e., horizontal, vertical, and adjacent) by rating on a 7-point Likert scale.

After the participant submitted the survey, his or her responses together with the survey questions would be shown on a new page. Then I would start the short interview by directing the

participant's attention to the survey results and asked three main questions: a) rationale for the ratings, b) process of allocating attention for the three presentation formats, and c) strategies for learning (see Appendix J for the interview questions). The interview generally lasted for about five minutes and was audio recorded. The audio recordings were transcribed into written texts for subsequent data analysis. Specifically, the transcriptions were first created with machine transcription (caption function in Kaltura MediaSpace), and then I listened to each interview and made revisions accordingly.

3.3.6 Background Survey

I used Qualtrics (www.qualtrics.com) to create an online background survey (see Supplementary Materials D for the print version of the complete survey). Participants would answer questions about their background, including age, gender, L1, L2, Chinese learning and study-abroad experience, self-rated Chinese proficiency level, vision deficiency, and learning difficulty.

3.4 Procedure

Table 15 presents the summary of procedure with estimated time. Each participant completed all activities on a computer individually. They started with the Chinese proficiency test and the pretest, then studied the target words and completed the posttest, followed by a post-learning survey and an interview. After that, they finished the working memory tasks and finally the background survey.

During the learning phase, participants would study the target words in groups, that is, they would study one group of words in one presentation format, and then would move to

another group of words in another presentation format. After studying all words for the first time, participants would study these words for a second time. During each time, the group order and the word order within each group were randomized for each participant. Table 16 shows the details of the three versions of the learning phase when presentation formats and word groups were counterbalanced according to a Latin square design. Each participant completed one version of the learning phase.

Table 15. Procedure

	Activity	Minutes
Proficiency	Chinese Proficiency Test	12
Pretest	Recognition and Recall Tasks	20
Learning	30 Two-character Words	20
Posttest	Recognition and Recall Tasks	35
Post-learning	Survey	2
	Interview	5
Working Memory	Visual Forward Digit Span	5
	Stroop	8
	Letter Memory	9
	Number Letter	2
Background	Survey	2

Table 16. Three Versions of Learning Phase

Version	Wordlist	a	b	c
1	i	H	V	A
2	ii	V	A	H
3	iii	A	H	V

Note. a = Group a. b = Group b. c = Group c.
A = Adjacent. H = Horizontal. V = Vertical.

To learn each word, each time participants would first see the word on the computer screen accompanied by its spoken pronunciation for two seconds (see Figure 5). Then, the word together with its pinyin and English translation would appear on the screen simultaneously and would stay for 15 seconds, during which time participants would study the word by viewing,

with their eye movements recorded. Using two 17-second sessions for a total of 34 seconds, which was close to Lee and Kalyuga's (2011) 30 seconds in total, was supported by the pilot data (see Appendix C for details) that participants demonstrated varying amount of lexical knowledge of the target words after studying. The time order of the two learning sessions will also serve as a variable in subsequent analysis of the eye-tracking data.

For the learning phase, I used SR Research Experiment Builder 2.1.140 (SR Research Ltd., 2017) to program an eye-tracking experiment. The characters were presented in 70 Microsoft YaHei font (black) and the pinyin and the English translation were in 35 Courier New font (bolded, black) on a grey background displayed by a 24' monitor with 1920*1080 resolution (ASUS VG248QE). Using Microsoft YaHei and Courier New fonts were because every print unit (i.e., a Chinese character or an alphabetic letter) had the same width within the same font style, which provided an ideal control for examining eye movements. The distance between the characters, pinyin, and English translation was 3.5-inch across all presentation formats. The eye-tracker were EyeLink 1000 (1000 Hz, SR Research Ltd.) with a desktop mount.

For each presentation format, three areas of interest (AOIs) were drawn for the Chinese characters, the pinyin, and the English translation separately (see Figure 6). All AOIs were of the same size, that is, 344*230 resolution.

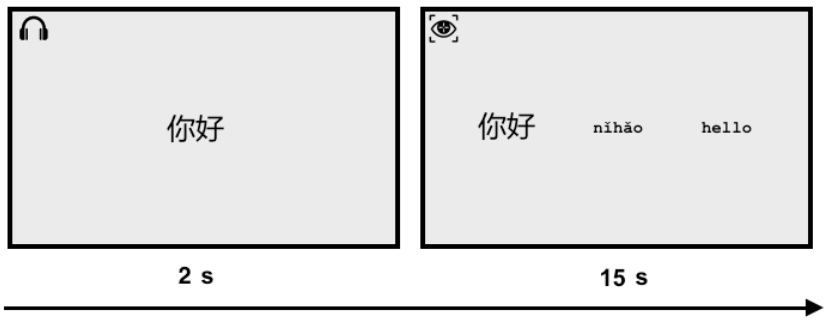

Figure 5. Learning phase.

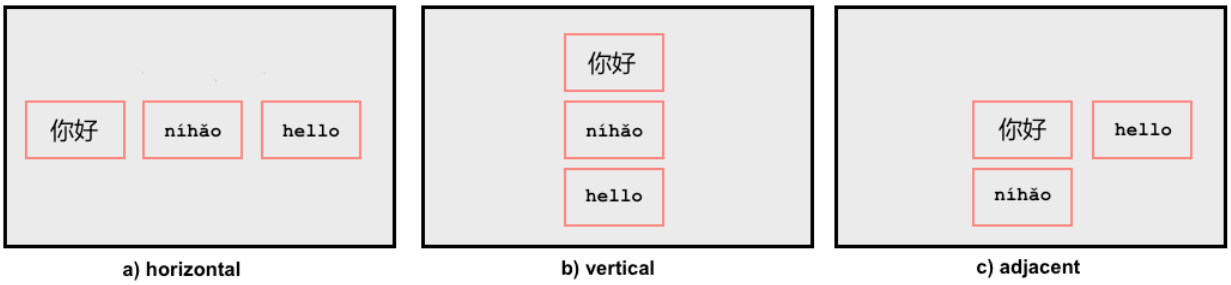

Figure 6. AOIs for three presentation formats.

3.5 Data Analysis

3.5.1 Data Indices

Table 17 summarizes the data indices generated from the instruments that measured different variables in the current study. Notably, for learning outcome, since the vocabulary pretest was expected to generate scores approaching 0 (i.e., participants had little prior knowledge of the target vocabulary), a gain score from pretest to posttest, instead of two separate scores for pretest and posttest, was calculated for each presentation format (i.e., horizontal, vertical, adjacent) for each participant.

Table 17. Variables, Instruments, and Data Indices

Variable	Instrument	Data Index
Learning Outcome	Vocabulary Pretest and Posttest	Pretest Score, Posttest Score, Gain Score
Learner Attention	Eye Tracking	Fixation Duration
		Fixation Count
Working Memory Capacity	Visual Forward Digit Span	Span
	Letter Memory	Accuracy Rate
	Number Letter	RT Difference
	Stroop	RT Difference
Learner Preference	Post-learning Survey	Rating
	Post-learning Interview	Verbal Report
Proficiency	Chinese Proficiency Test	Score

Table 18. Sub-score Indices, Test Items, and Tasks

Index	Test Item*	Task
M2C_rcl	(1) From Meaning to Characters	Recall
M2P_rcl	(2) From Meaning to Pinyin	Recall
M2C_rcg	(3) From Meaning to Characters	Recognition
M2P_rcg	(4) From Meaning to Pinyin	Recognition
C2M_rcl	(5) From Characters to Meaning	Recall
C2P_rcl	(6) From Characters to Pinyin	Recall
C2M_rcg	(7) From Characters to Meaning	Recognition
C2P_rcg	(8) From Characters to Pinyin	Recognition
M2C	(1) From Meaning to Characters	Recall
	(3) From Meaning to Characters	Recognition
M2P	(2) From Meaning to Pinyin	Recall
	(4) From Meaning to Pinyin	Recognition
C2M	(5) From Characters to Meaning	Recall
	(7) From Characters to Meaning	Recognition
C2P	(6) From Characters to Pinyin	Recall
	(8) From Characters to Pinyin	Recognition
RCL	(1) From Meaning to Characters (2) From Meaning to Pinyin (5) From Characters to Meaning (6) From Characters to Pinyin	Recall
RCG	(3) From Meaning to Characters (4) From Meaning to Pinyin (7) From Characters to Meaning (8) From Characters to Pinyin	Recognition

Note. * For details of the test items, see Table 12.

Apart from these overall indices, sub-scores were also calculated for each of the eight test formats, for each of the four mappings, and for each of the two test formats (see Table 18) as recommended, because different test formats may represent different extents of strength of vocabulary knowledge (Laufer et al., 2004; Nation, 2013).

3.5.2 Overall Analytical Approach and Statistical Methods

In language studies, participants are usually recruited as a sample from the population, and they are often treated as a random effect in statistical analysis (Baayen, Davidson, & Bates, 2008; Cunnings, 2012; Linck & Cunnings, 2015). Contrarily, language stimuli are less often included as a random effect in analysis, and the "language-as-fixed-effect-fallacy" (Clark, 1973) argues that both participants and language stimuli should be accounted for as random effects at the same time in a single analysis, because language stimuli are also randomly sampled from indefinite possibilities (Baayen et al., 2008; Cunnings, 2012; Cunnings & Finlayson, 2015; Linck & Cunnings, 2015). Mixed effects modeling enabled by modern computational technology has offered an effective analytical solution to include both participants and language stimuli as random effects (Baayen et al., 2008), and L2 researchers (e.g., Cunnings, 2012; Cunnings & Finlayson, 2015; Linck & Cunnings, 2015) have been advocating the use of mixed effects modeling as a way to advance the field statistically. Specifically, the random effects in mixed effects models can be nested/hierarchical or crossed. For nested/hierarchical random effects, participants can be students who come from different classes of the same school, and the class unit can add another layer of randomness to each student's individuality, with students nested within classes (Cunnings & Finlayson, 2015; Linck & Cunnings, 2015). For crossed random effects, all participants can have experienced all language stimuli presented in different conditions, with participants crossed with language stimuli (Baayen et al., 2008). As mentioned in 3.1 Overall Design and Operationalization, the current study adopted a within-subject design and every participant experienced each language stimulus, with multiple data points from the same participant (i.e., repeated measurement data, Baayen et al., 2008). Therefore, the current study fell within the crossed random effects category.

To address the idea of "language-as-fixed-effect-fallacy" (Clark, 1973), this study adopted mixed effects modeling as the major quantitative analytical approach to include both participants and language stimuli as random effects. Notably, mixed effects models can be run directly on raw data that are usually at item level and generally do not require data aggregation at condition/participant level, that is, calculating (sub-)sum scores for each condition/participant (Cunnings, 2012; Linck & Cunnings, 2015). Depending on the distribution of the data of the outcome variable, a particular mixed effects modeling method was selected (Cunnings & Finlayson, 2015; Zuur, Ieno, Walker, Saveliev, & Smith, 2009) for quantitative analysis in this study.

As Larson-Hall and Plonsky (2015) emphasized, descriptive statistics are "absolutely necessary" and "fundamentally essential" in quantitative analysis (p. 130). Following their advice, I calculated descriptive and normality statistics for all quantitative data indices. To explore the potential relationships between the quantitative data indices, I first performed bootstrapped Pearson's correlations for all data indices (Field, 2018; Larson-Hall, 2016). Then I turned to each research question (except for RQ 4) and conducted mixed effects modeling, as well as repeated-measures ANOVA (for RQ 4). Participants' verbal reports were drawn on from time to time to provide supplementary information to the quantitative results. Detailed statistical methods will be described in Chapter 4 Analysis and Results so as to provide a clear account of how the results were generated to answer the research questions.

CHAPTER 4 ANALYSIS AND RESULTS

Given the importance of descriptive statistics in quantitative analysis of L2 research (Larson-Hall & Plonsky, 2015), I will first present the analysis and results of descriptive statistics and normality tests for each quantitative data index (see Tables 17 & 18), followed by the analysis and results of bivariate correlations to explore the potential relationships between the quantitative data indices. Then I will report the analysis and results for each RQ.

4.1 Descriptive Statistics and Normality Tests

4.1.1 Vocabulary Pretest and Posttest

For the data of the vocabulary score indices (see Tables 17 & 18), descriptive statistics of mean, standard deviation (SD), minimum (Min), and maximum (Max) were calculated for each presentation format using IBM SPSS Statistics 25 (see Table 19). Notably, because of the current within-subject design, the confidence intervals (CIs) of the mean need to be calculated with adjustments to the common method for calculating between-subjects CIs (see Loftus & Masson, 1994 for discussion on the differences between within- and between-subjects CIs). Building on Loftus and Masson's (1994) seminal proposal, several refined methods for within-subject CI calculation and plotting have been proposed (e.g., Cousineau, 2005; Franz & Loftus, 2012; Goldstein & Healy, 1995; Morey, 2008; see Baguley, 2012 for review of these methods). For the purpose of exploring the pattern among means as well as mean differences in a moderate to large sample, Baguley (2012) recommended using the Cousineau-Morey method with adjustments and offered ready-to-use R functions for calculating and plotting within-subject CIs, which is confirmed by other researchers as appropriate (e.g., Cousineau & O'Brien, 2014). Given my purpose of comparing the effects of three presentation formats on vocabulary gain scores, I

followed Baguley (2012) in adopting the Cousineau-Morey method and using the cm.ci function to calculate the 95% CIs (i.e., C-M 95% CI of Mean, see Table 19) for all vocabulary score indices, using R 4.0.2 via RStudio 1.3.1056. In addition, I used the plot.wsci function by Baguley (2012) to generate individual CI plots for each vocabulary score index (see Figures 7, 8, & 9). (For the complete R codes described in Baguley, 2012, see https://osf.io/6768q/?show=view)

To assess whether the data were normally distributed, statistical tests were performed to check whether the values of skewness and kurtosis were significantly different from 0 (see Figure 7 for the equation). If the absolute value of z is larger than 1.96, the test result is significant at $p < .05$. If it is larger than 2.58, the test result is significant at $p < .01$ (Field, 2018; Hair et al., 2019). Given the small to moderate sample size in this study, $p < .01$ with the absolute value of z larger than 2.58 is recommended as the significance level (Tabachnick & Fidell, 2018). As shown in Table 19, most of the gain scores were normally distributed.

$$z = \frac{Skewness/Kurtosis}{Standard\ Error\ of\ Skewness/Kurtosis}$$

Figure 7. Equation for calculating the z score for skewness and kurtosis. Adapted from Field (2018), Hair et al. (2019, pp. 95-96), and Tabachnick and Fidell (2018, p. 69).

As in Table 19, when comparing the means of vocabulary gain scores among the three presentation formats, the adjacent format was associated with the highest gain score for most of the vocabulary test formats, except for M2P (vertical the highest), M2P_rcg (vertical the highest), and C2P_rcl (horizontal the highest). When considering the C-M 95% CI plots in Figures 8, 9, and 10, the vocabulary gain score of the adjacent format may be significantly higher than those of the other two formats in cases where all vocabulary test formats are included (see panel All in Figure 8) and where both recall and recognition tasks are included in C2M (see panel C2M in Figure 8), as indicated by little overlap between the whiskers. Notably, when recall

and recognition tasks are considered separately for C2M (see panel C2M_rcl in Figure 9 and panel C2M_rcg in Figure 10), there is not much overlap either, indicating there might be statistically significant differences from the score of the adjacent format to those of the other two formats. These results implied that when vocabulary gain scores were compared at the group level without considering the effects of other factors, such as L2 Chinese proficiency, the adjacent format was associated with an overall higher score than the other two formats, and that the differences may mainly come from the C2M recall and recognition tasks, that is, going from L2 form to L1 meaning.

Table 19. Descriptive and Normality Statistics for Vocabulary Gain Scores

		Mean	*C-M 95% CI of Mean*		*SD*	*Min*	*Max*	$z_{Skewness}$	$z_{Kurtosis}$
	H	20.12	[19.41,	20.82]	10.46	1	42.67	0.58	-1.2
All	V	20.12	[19.41,	20.83]	9.85	2	43	0.67	-0.98
	A	21.49	[20.76,	22.21]	10.38	1	44.67	0.57	-1.31
	H	2.93	[2.68,	3.17]	2.96	0	11.50	3.13*	0.07
RCL	V	2.67	[2.44,	2.90]	3.04	0	12	3.68*	0.62
	A	3.20	[2.95,	3.45]	3.16	0	11.50	2.17	-1.29
	H	17.19	[16.58,	17.79]	8.41	1	36	0.49	-0.95
RCG	V	17.45	[16.87,	18.02]	7.73	2	34	0.10	-0.82
	A	18.29	[17.69,	18.89]	8.24	1	35	0.40	-1.10
		Mean	*C-M 95% CI of Mean*		*SD*	*Min*	*Max*	$z_{Skewness}$	$z_{Kurtosis}$
	H	5.40	[5.18,	5.61]	2.93	0	10	-0.31	-1.67
M2C	V	5.44	[5.22,	5.65]	2.74	0	10	-1.49	-1.47
	A	5.71	[5.47,	5.95]	2.89	0	10.5	-1.02	-1.53
	H	3.32	[3.10,	3.54]	2.34	0	9	2.67*	-0.07
M2P	V	3.68	[3.45,	3.91]	2.09	0	8.33	1.98	-0.76
	A	3.59	[3.38,	3.80]	2.10	0	9	1.80	-0.06
	H	8.01	[7.70,	8.33]	5.03	0	20	1.16	-1.10
C2M	V	7.67	[7.32,	8.01]	4.94	0	19	1.08	-1.03
	A	8.59	[8.26,	8.92]	5.10	0	20	0.59	-1.45
	H	3.38	[3.13,	3.63]	2.76	0	12	2.74*	0.54*
C2P	V	3.34	[3.11,	3.56]	2.52	0	10.67	3.04*	0.64
	A	3.59	[3.34,	3.84]	3.00	0	11.50	2.69*	-0.19
		Mean	*C-M 95% CI of Mean*		*SD*	*Min*	*Max*	$z_{Skewness}$	$z_{Kurtosis}$
	H	0.01	[-0.01,	0.03]	0.06	0	0.5	28.74*	121.05*
M2C_rcl	V	0.03	[0.01,	0.05]	0.15	0	1	19.06*	56.38*
	A	0.04	[0.02,	0.06]	0.17	0	1	14.37*	32.19*

Table 19 (cont'd)

	H	5.39	[5.17,	5.61]	2.93	0	10	-0.29	-1.69
M2C_rcg	V	5.41	[5.19,	5.63]	2.73	0	10	-1.42	-1.45
	A	5.67	[5.42,	5.91]	2.86	0	10	-0.30	-0.90
	H	0.05	[0.02,	0.08]	0.18	0	1	14.34*	30.20*
M2P_rcl	V	0.06	[0.02,	0.10]	0.30	0	2.33	23.92*	90.70*
	A	0.06	[0.04,	0.07]	0.19	0	0.83	11.63*	17.89*
	H	3.28	[3.05,	3.50]	2.32	0	9	2.86*	0.16
M2P_rcg	V	3.62	[3.39,	3.85]	2.04	0	8	1.88	-0.81
	A	3.54	[3.33,	3.75]	2.06	0	9	1.82	-0.02
		Mean	*C-M 95% CI of Mean*		*SD*	*Min*	*Max*	$z_{Skewness}$	$z_{Kurtosis}$
	H	2.28	[2.09,	2.47]	2.47	0	10	3.88*	0.93
C2M_rcl	V	2.12	[1.93,	2.30]	2.50	0	10	4.00*	0.96
	A	2.55	[2.35,	2.75]	2.61	0	10	2.63*	-0.65
	H	5.74	[5.55,	5.93]	3.10	0	10	-0.94	-1.80
C2M_rcg	V	5.55	[5.34,	5.76]	2.98	0	10	-1.09	-1.43
	A	6.04	[5.84,	6.25]	3.12	0	10	-1.37	-1.84
	H	0.60	[0.51,	0.68]	0.85	0	4	5.80*	5.47*
C2P_rcl	V	0.47	[0.39,	0.55]	0.82	0	4	8.48*	12.30*
	A	0.55	[0.46,	0.64]	0.80	0	3.17	5.55*	3.43*
	H	2.78	[2.57,	3.00]	2.26	0	9	2.45	-0.14
C2P_rcg	V	2.87	[2.66,	3.08]	2.00	0	8	2.21	-0.23
	A	3.04	[2.81,	3.27]	2.51	0	9	2.26	-0.96

Note. $n = 69$. C-M = Cousineau-Morey. H = Horizontal. V = Vertical. A = Adjacent. The highest score among the three presentation formats for each test format has been highlighted.
*$p < .01$.

4.1.2 Eye-tracking

Due to technical problem, one participant's eye movement data was lost, thus leaving 68 participants for analysis. Descriptive statistics of mean, standard deviation (SD), minimum (Min), and maximum (Max) were calculated for each element in each presentation format using IBM SPSS Statistics 25. Given the current within-subject design, the Cousineau-Morey method with adjustments (Baguley, 2012) was used to calculate C-M 95% CIs of the mean and to create C-M 95% CI plots, using R 4.0.2 via RStudio 1.3.1056. To assess data normality, *z* scores of skewness and kurtosis were also calculated (see Figure 7 for the equation). Tables 20-21 present

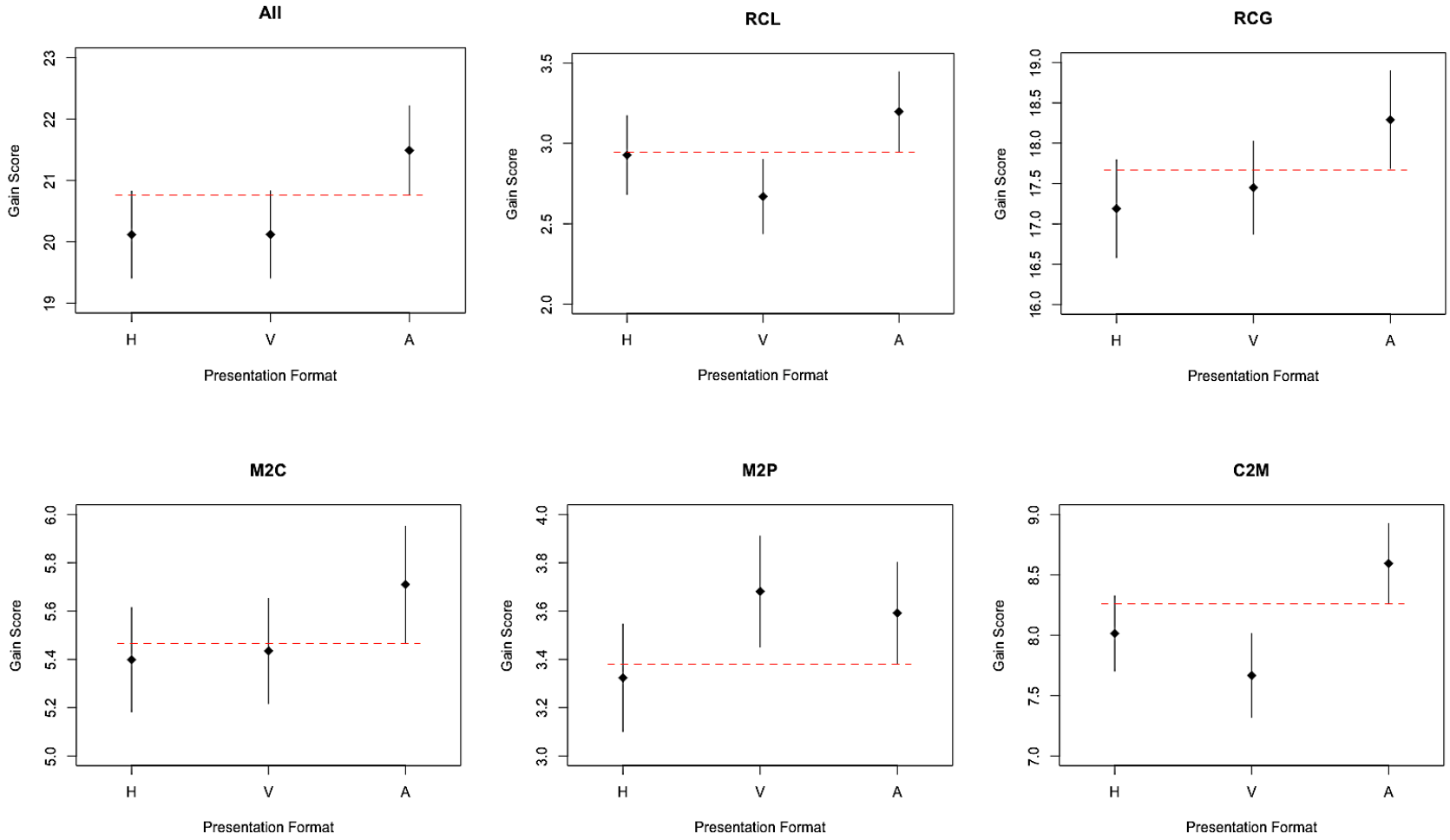

Figure 8. C-M 95% CI plots for each vocabulary score index (All, RCL, RCG, M2C, M2P, C2M). A red dash-line was drawn at the lower bound of the adjacent format.

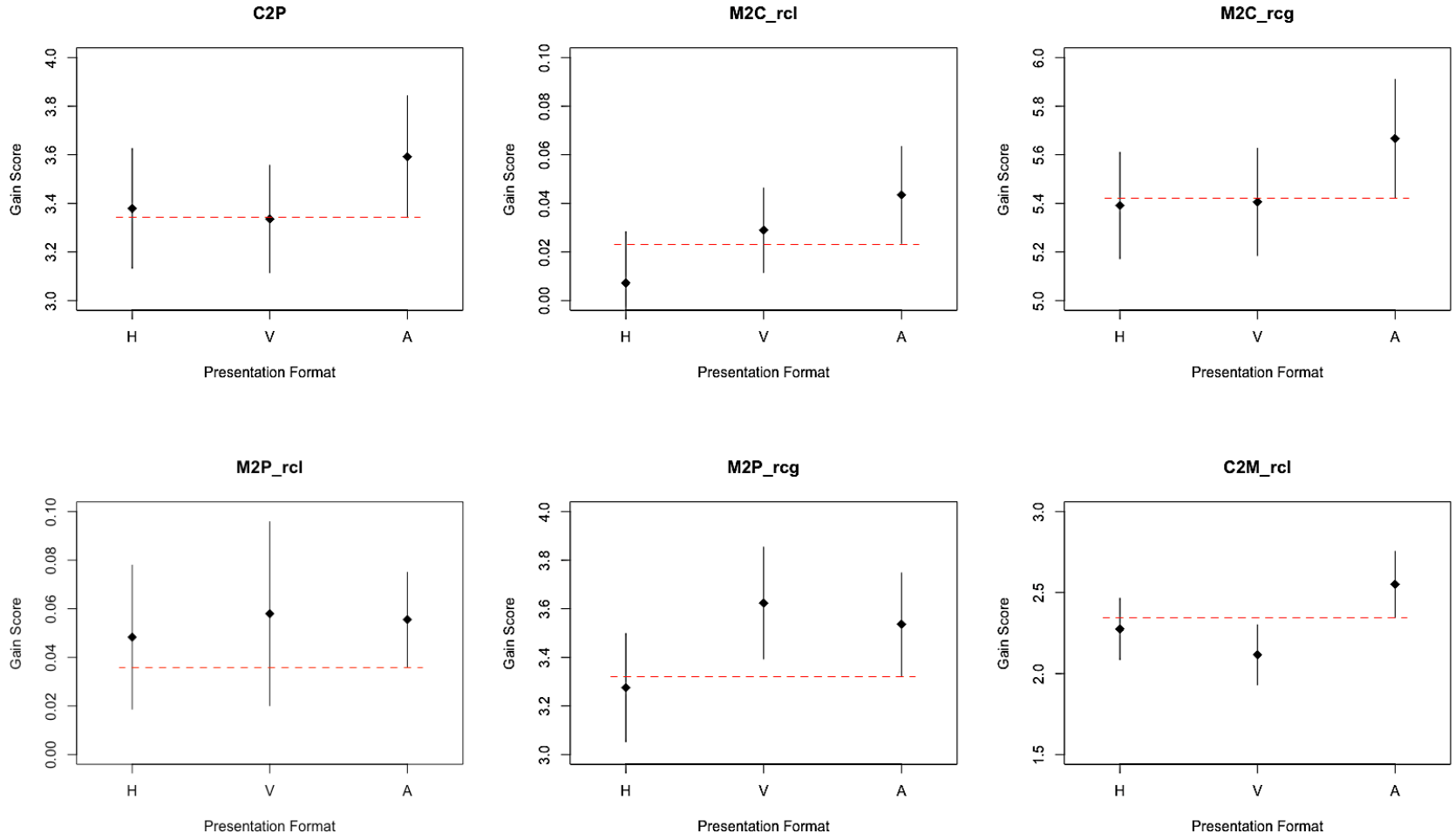

Figure 9. C-M 95% CI plots for each vocabulary score index (C2P, M2C_rcl, M2C_rcg, M2P_rcl, M2P_rcg, C2M_rcl). A red dash-line was drawn at the lower bound of the adjacent format.

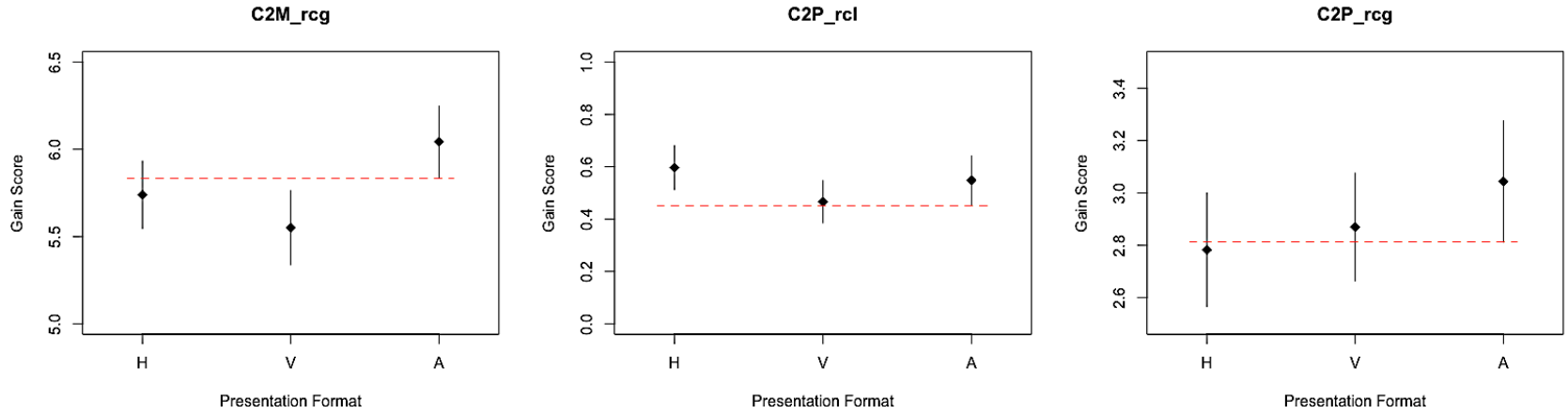

Figure 10. C-M 95% CI plots for each vocabulary score index (C2M_rcg, C2P_rcl, C2P_rcg). A red dash-line was drawn at the lower bound of the adjacent format.

Table 20. Descriptive and Normality Statistics for Fixation Durations in Three Presentation Formats

		Mean	*C-M 95% CI of Mean*		*SD*	*Min*	*Max*	$z_{Skewness}$	$z_{Kurtosis}$
H	Character	7918.87	[7595.20,	8241.37]	1955.26	2899.00	11655.45	-1.32	-0.53
	Pinyin	3135.60	[2903.79,	3367.40]	1181.34	765.20	6001.85	1.92	0.06
	Meaning	1485.61	[1362.82,	1608.39]	600.34	496.05	3460.70	2.82*	1.39
V	Character	8089.39	[7743.63,	8435.15]	2041.41	588.05	11505.25	-2.72*	2.34
	Pinyin	3104.83	[2853.95,	3355.71]	1321.53	1015.15	8564.25	4.68*	5.80*
	Meaning	1270.83	[1161.72,	1379.94]	508.31	513.05	2748.20	3.01*	0.46
A	Character	8787.17	[8512.84,	9061.51]	1668.40	1693.65	11845.90	-3.52*	6.12*
	Pinyin	2222.87	[2054.78,	2390.96]	887.56	869.75	5694.95	4.33*	5.03*
	Meaning	1564.13	[1412.11,	1716.15]	735.84	389.40	4527.95	5.13*	5.74*

Note. $n = 68$. H = Horizontal. V = Vertical. A = Adjacent.
*$p < .01$.

Table 21. Descriptive and Normality Statistics for Fixation Counts in Three Presentation Formats

		Mean	*C-M 95% CI of Mean*		*SD*	*Min*	*Max*	$z_{Skewness}$	$z_{Kurtosis}$
H	Character	20.36	[19.58,	21.14]	5.01	9.35	31.9	0.04	-0.37
	Pinyin	10.64	[10.07,	11.22]	3.46	3.60	22.85	2.09	1.93
	Meaning	4.80	[4.47,	5.12]	1.59	2.00	8.70	1.59	-0.30
V	Character	20.31	[19.43,	21.19]	5.57	2.65	33.60	-0.60	1.27
	Pinyin	10.94	[10.32,	11.56]	3.72	4.85	23.75	3.45*	2.71
	Meaning	4.69	[4.37,	5.01]	2.46	1.80	8.25	2.35	-0.22
A	Character	22.28	[21.59,	22.97]	4.78	5.50	31.40	-2.37	2.60*
	Pinyin	8.37	[7.91,	8.84]	2.73	3.65	18.25	3.63*	3.11*
	Meaning	5.33	[4.94,	5.71]	2.05	1.85	13.20	4.38*	4.63*

Note. n = 68. H = Horizontal. V = Vertical. A = Adjacent.
*$p < .01$.

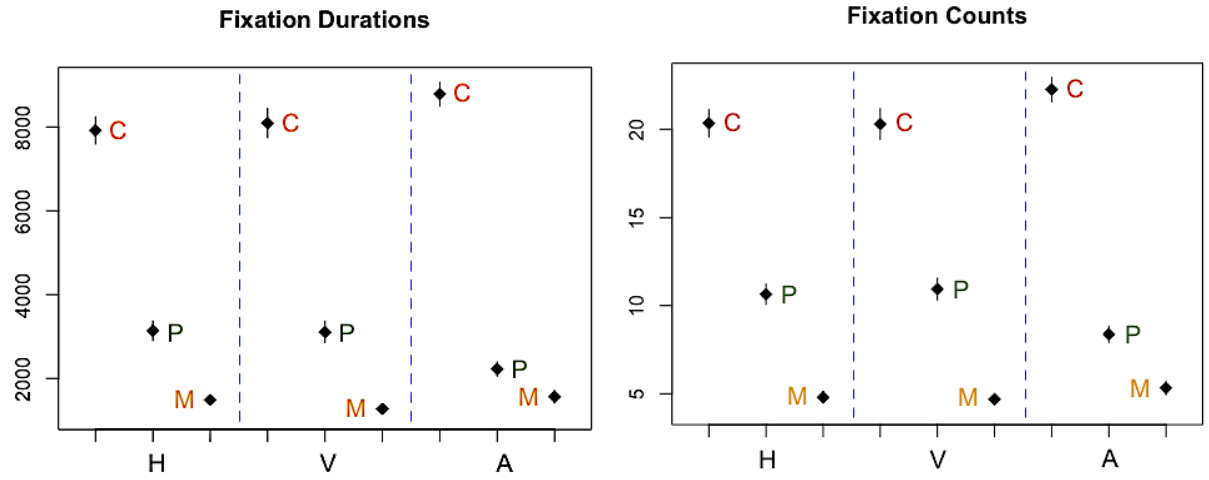

Figure 11. C-M 95% CI plots for fixation durations and fixation counts. Blue dash-lines were drawn to separate different presentation formats.

the results of descriptive and normality statistics for fixation durations and fixation counts on characters, pinyin, and meaning. Based on the criterion of $p < .01$ with the absolute value of z larger than 2.58 (Tabachnick & Fidell, 2018), some of the eye-movement data were not normally distributed. Figure 11 shows the corresponding C-M 95% CI plots.

As shown in Figure 11, the eye-movement patterns of the horizontal and vertical formats were very similar to each other, but very different from that of the adjacent format. Compared

with the horizontal and vertical formats, the adjacent format received significantly longer fixation durations and more fixation counts to characters and meaning, but significantly less of those to pinyin (as indicated by little overlap of the whiskers). Within each format, characters received significantly the most fixation durations and fixation counts, whereas meaning received significantly the least of them (as indicated by large spaces between the whiskers).

4.1.3 Working Memory Tasks

Visual Forward Digit Span and Letter Memory Tasks. Bootstrapped descriptive statistics were calculated for the data of the two tasks based on 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. Notably, due to personal reason, one participant did not complete the letter memory task. Following Friedman et al. (2016), between-subjects data trimming was performed to reduce the influence of extreme data points. Specifically, data points that were beyond 3 *SD*s from the group mean were replaced by values 3 *SD*s from the mean. No data was affected during the between-subjects data trimming.

Table 22. Bootstrapped Descriptive and Normality Statistics for Four Working Memory Tasks

		Digit Span	Letter Memory	Number Letter	Stroop
n		69	68	60	65
Mean		6.74	0.71	773.23	152.49
BCa 95%	*Lower*	6.52	0.68	680.52	132.29
CI of Mean	*Upper*	6.97	0.74	871.98	173.41
SD		1.05	0.11	374.12	83.08
Min		5	0.44	185.12	25.69
Max		9	0.94	1857.74	404.14
$z_{Skewness}$		2.43	-0.53	3.62*	2.05
$z_{Kurtosis}$		-0.32	-0.36	1.85	-0.05

Note. *$p < .01$.

Table 22 presents the descriptive statistics for the two tasks. *Z* scores of skewness and kurtosis were calculated to assess data normality (see Figure 7 for the equation). According to

the significance level of $p < .01$ with the absolute value of z larger than 2.58 (Tabachnick & Fidell, 2018), the data of both tasks were normally distributed.

Number Letter Task. Data screening and trimming were conducted for the data of this task by following Friedman et al. (2008, 2016), using RStudio 1.1.447. Specifically, data from 9 participants were discarded due to their accuracy rates lower than 92%, leaving 60 participants for further analysis. Then, within-subject data trimming was performed for each participant's RT of each trial. Specifically, RTs of error trials, trials that were less than 200 ms, and trials that immediately followed error trials were excluded. As mentioned previously, task performance was indexed by RT differences between target trials that required mental shifting and the control trials that did not. Following Friedman et al. (2008, 2016), in order to provide optimal measures of the central tendency of the data for target and control trials, I adopted Wilcox and Keselman's (2003) robust method of excluding outliers by calculating the median and the median absolute deviation (MAD). Specifically, for each condition (i.e., target, control), RTs that were away from the median by more than 3.32 times the MAD (see Figure 12 for the equation) were excluded, resulting in 15.31% of the data discarded. This trimming rate may not be alarming, because higher trimming rates have been recommended by statisticians, such as 20% by Wilcox and Keselman (2003, p. 267). Based on the trimmed RTs, the mean was calculated for each condition (i.e., target, control) for each participant. The RT difference was obtained by subtracting the mean of the control condition from the mean of the target condition. The interpretation of the RT differences should be inverse: the smaller the value, the better the task performance. Then, between-subjects data trimming was performed to reduce the influence of extreme RTs (Friedman et al., 2016). Specifically, RTs that were beyond 3 *SD*s from the group mean were replaced by values 3 *SD*s from the mean. No data was affected during the between-subjects data

trimming.

$$\frac{|RT - Median|}{MAD} > 3.32$$

Figure 12. Equation for spotting RT outliers. Adapted from Wilcox and Keselman (2003, p. 264).

Table 22 presents the bootstrapped descriptive statistics of the data based on 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. To assess data normality, *z* scores of skewness and kurtosis were calculated (see Figure 7 for the equation). Based on the significance level of $p < .01$ with the absolute value of *z* larger than 2.58 (Tabachnick & Fidell, 2018), the data of the number letter task were not normally distributed.

Stroop Task. As mentioned in 3.2 Participant, two participants self-reported that they had colorblindness, so they were excluded from the data set of this task. Similar to the number letter task, data screening and trimming were conducted for the data by following Friedman et al. (2008, 2016), using RStudio 1.1.447. Specifically, data from two participants were discarded due to their accuracy rates lower than 92%, leaving 65 participants for further analysis. Then, within-subject data trimming was performed for each participant's RT of each trial in the incongruent condition that required mental inhibition and the control condition that did not. Specifically, RTs of error trials and trials that were less than 200 ms were excluded. Similar to the number letter task, performance of the Stroop task was indexed by RT differences between incongruent trials and control trials. Following Friedman et al. (2008, 2016), I adopted Wilcox and Keselman's (2003) robust method of excluding outliers in the same manner as I did for the number letter task. The trimming procedure resulted in 11.27% of the data discarded, which is acceptable as mentioned previously. Then, between-subjects data trimming was performed to reduce the influence of extreme RTs (Friedman et al., 2016). Specifically, RTs that were beyond 3 *SD*s

from the group mean were replaced by values 3 *SD*s from the mean. One participant's data was trimmed during the between-subjects data trimming.

Table 22 presents the bootstrapped descriptive statistics of the data based on 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. *Z* scores of skewness and kurtosis were calculated to assess data normality (see Figure 7 for the equation). According to the significance level of $p < .01$ with the absolute value of z larger than 2.58 (Tabachnick & Fidell, 2018), the data was normally distributed.

Composite Scores from Principal Component Analysis. Following Indrarathne and Kormos (2017), I conducted principal component analysis on the data of the working memory tasks to test Miyake and colleagues' unity/diversity framework (Miyake & Friedman, 2012; Miyake, Friedman et al., 2000). According to this framework, the three executive functions, namely, updating, shifting, and inhibition, are correlated yet separable.

As shown in Table 22, except the digit span task ($n = 69$), the other three tasks had missing values to different extents: 1.45% for letter memory ($n = 68$), 13.04% for number letter ($n = 60$), and 5.80% for Stroop ($n = 65$). Following the missing data literature's recommendation of imputing rather than deleting the missing values (Enders, 2010; Little & Rubin, 2020; Raghunathan, 2015), I used R 3.1.3 via RStudio 1.1.447 with the missMDA (version 1.7.3) package (Josse & Husson, 2016) to impute the missing values in the working memory tasks. (see Appendix K for detailed discussion of the rationale.)

Table 23 presents the bootstrapped descriptive statistics of the imputed dataset based on 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. *Z* scores of skewness and kurtosis were calculated to assess data normality (see Figure 7 for the equation). According to the significance level of $p < .01$ with the absolute value

of *z* larger than 2.58 (Tabachnick & Fidell, 2018), except the number letter task, other tasks had normally distributed data. Since principal component analysis assumes multivariate normality, square root transformation is recommended to remedy moderately nonnormal distributions (Tabachnick & Fidell, 2018). Therefore, square roots were calculated for the data of all tasks. Table 24 presents the bootstrapped descriptive statistics of the square rooted, imputed dataset based on 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016). After the transformation, all data were normally distributed. Following Friedman et al. (2008), all RT differences were reversed so that for all working memory measures, the larger the value, the better the task performance.

As suggested by Tabachnick and Fidell (2018), multivariate outliers were identified based on the criterion of Mahalanobis distance at $p < .001$, using IBM SPSS Statistics 25. Specifically, Mahalanobis distance was checked through chi-square, with degrees of freedom equal to number of variables (Tabachnick & Fidell, 2018). For the data of the four working memory tasks, the degree of freedom was 4, and the case with Mahalanobis distance larger than 18.467 would be identified as a multivariate outlier. With the current dataset, the largest Mahalanobis distance was 10.214, so no multivariate outlier was found.

Table 23. Bootstrapped Descriptive and Normality Statistics for Imputed Data of Working Memory Tasks

		Digit Span	Letter Memory	Number Letter	Stroop
Mean		6.74	0.71	772.42	152.85
BCa 95% CI of Mean	*Lower*	6.52	0.68	688.30	133.45
	Upper	6.97	0.74	856.54	172.25
SD		1.05	0.11	350.18	80.76
Min		5	0.44	185.12	25.69
Max		9	0.94	1857.74	404.14
$z_{Skewness}$		2.43	-0.44	4.11*	2.11
$z_{Kurtosis}$		-0.32	-0.36	2.89*	0.21

Note. *$p < .01$. $n = 69$.

Table 24. Bootstrapped Descriptive and Normality Statistics for Imputed, Square-rooted Data of Working Memory Tasks

		DgtSpn_sqrt	LetMem_sqrt	NumLet_sqrt	Strp_sqrt
Mean		2.59	0.84	27.14	11.91
BCa 95%	*Lower*	2.54	0.82	25.69	11.10
CI of Mean	*Upper*	2.64	0.85	28.59	12.72
SD		0.20	0.07	6.04	3.36
Min		2.24	0.66	13.61	5.07
Max		3.00	0.97	43.10	20.10
$z_{Skewness}$		1.92	-1.19	1.78	-0.12
$z_{Kurtosis}$		-0.55	-0.03	0.88	-0.76

Note. n = 69. DgtSpn = Digit Span (DS). LetMem = Letter Memory (LM). NumLet = Number Letter (NL). Strp = Stroop (ST).
$^{*}p < .01$.

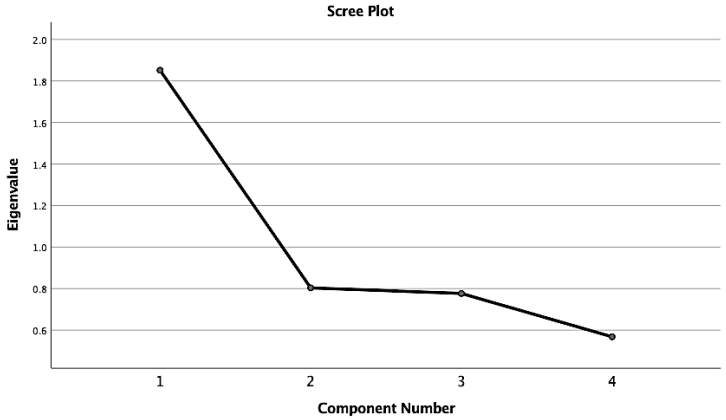

Figure 13. Scree plot of principal component analysis.

Principal component analysis was performed on the dataset with IBM SPSS Statistics 25, with direct oblimin (for correlated variables, Field, 2018) as the rotation method. Multi-collinearity was not found to pose any problem, as the determinant of the correlation matrix for all variables was 0.657, above the 0.00001 threshold (Field, 2018). Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was 0.678, and which was larger than the recommended value of 0.5 (Kaiser & Rice, 1974) and indicated the current sample size was adequate for conducting principal component analysis. Bartlett's test of specificity was statistically significant at $p < .001$,

which supported the correlation matrix was factorable (Field, 2018). Based on Kaiser's (1960, 1970) criterion of retaining factors with eigenvalues larger than 1, one component was extracted: eigenvalue = 1.852, explaining 46.29% variance. Scree plot (see Figure 13) also confirmed that one component could be extracted from the four variables. The component loading was 0.685 for DgtSpn_sqrt, 0.779 for LetMem_sqrt, 0.606 for NumLet_sqrt, and 0.638 for Strp_sqrt.

To assess the correlations between the four variables, bootstrapped Pearson's correlations were performed on 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. As shown in Table 25, LetMem_sqrt correlated with DgtSpn_sqrt, NumLet_sqrt, and Strp_sqrt at $p < .05$. In addition, the BCa 95% CI of r did not cross zero for the correlations between DgtSpn_sqrt and NumLet_sqrt, and between NumLet_sqrt and Strp_sqrt, indicating that these two correlations were also statistically significant (Field, 2018). The only nonsignificant correlation was between DgtSpn_sqrt and Strp_sqrt. According to Plonsky and Oswald's (2014) benchmarks of effect size ($r = .25$, small; $r = .40$, medium; $r = .60$, large), the significant correlations had small to medium effect size.

Table 25. Bootstrapped Pearson's Correlations for Imputed, Square-rooted Data of Working Memory Tasks

		LetMem_sqrt	NumLet_sqrt	Strp_sqrt
r	DgtSpn_sqrt	.40*	.22	.22
p		.001	.075	.071
BCa 95% CI of r		[.20, .58]	[.00, .41]	[-.02, .44]
r	LetMem_sqrt		.29*	.33*
p			.016	.005
BCa 95% CI of r			[.08, .48]	[.12, .53]
r	NumLet_sqrt			.22
p				.065
BCa 95% CI of r				[.02, .41]

Note. $n = 69$. DgtSpn = Digit Span (DS). LetMem = Letter Memory (LM). NumLet = Number Letter (NL). Strp = Stroop (ST).
*$p < .05$.

Table 26. Results of Principal Component Analysis for Imputed and Unimputed, Square-rooted Data of Working Memory Tasks

		Imputed	Unimputed
n		69	57
Determinant of Correlation Matrix		.657	.644
KMO of Sampling Adequacy		.678	.683
Bartlett's Test of Specificity		< .001	.001
Component	*Eigenvalue*	1.852	1.871
	% of Variance	46.29	46.79
Component Loading	DgtSpn_sqrt	.685	.711
	LetMem_sqrt	.779	.776
	NumLet_sqrt	.606	.581
	Strp_sqrt	.638	.653

Note. DgtSpn = Digit Span (DS). LetMem = Letter Memory (LM). NumLet = Number Letter (NL). Strp = Stroop (ST).

Table 27. Bootstrapped Pearson's Correlations for Unimputed, Square-rooted Data of Working Memory Tasks

		LetMem_sqrt	NumLet_sqrt	Strp_sqrt
r	DgtSpn_sqrt	.42*	.23	.25
p		.001	.086	.064
BCa 95% CI of r		[.20, .61]	[.01, .42]	[-.01, .49]
r	LetMem_sqrt		.26*	.35*
p			.049	.008
BCa 95% CI of r			[.03, .46]	[.12, .55]
r	NumLet_sqrt			.22
p				.109
BCa 95% CI of r				[.00, .41]

Note. $n = 57$. DgtSpn = Digit Span (DS). LetMem = Letter Memory (LM). NumLet = Number Letter (NL). Strp = Stroop (ST).
*$p < .05$.

To ensure that the current results were not caused by the data imputation process, the same principal component analysis with oblimin rotation (listwise deletion of missing data) and bootstrapped Pearson's correlation (pairwise deletion of missing data) were performed on the unimputed data of DgtSpn_sqrt, LetMem_sqrt, NumLet_sqrt, and Strp_sqrt. Still, one

component was extracted from the four variables with similar component loadings (see Table 26 to compare the results of imputed and unimputed data). For correlations, based on the *p* value and BCa 95% *CI* of *r*, similar pattern with similar effect size were obtained: except the correlations between DgtSpn_sqrt and Strp_sqrt and between NumLet_sqrt and Strp_sqrt, others were statistically significant with small to medium effect size (see Table 27). 95% CI plots were drawn for the correlation coefficient r for the imputed and unimputed data, using R 4.0.2 via RStudio 1.3.1056 (see Figure 14). Overall, results of the imputed data were comparable to those of the unimputed data, and thus may not be significantly affected by the data imputation process.

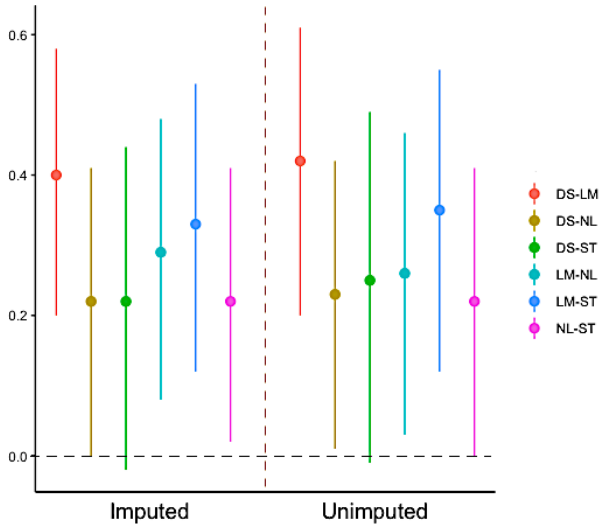

Figure 14. 95% CI plot for Pearson's correlations between the imputed and unimputed, square-rotted data of four working memory tasks. Brown dash-lines were drawn to separate different presentation formats, and a black dash-line was drawn at $r = 0$.

The current results were different from those in Indrarathne and Kormos (2017), which measured the same set of storage component and executive functions of working memory but used different tasks. Specifically, one component was extracted from the four variables in this study, whereas two components were extracted in Indrarathne and Kormos (2017). However, the current finding was not unexpected. Actually, Miyake and Friedman (2012) noted that one single

unitary component could be extracted from different working memory tasks, such as in Wiebe, Espy, and Charak's (2008) study. In addition, the sizes of significant correlations between the four working memory tasks in the current study were a bit smaller than those reported in Indrarathne and Kormos (2017) (which were above .50). This difference may be caused by the different task languages: L1 English was used in the current study, whereas L2 English was used in Indrarathne and Kormos (2017). Previous empirical studies (e.g., Gass & Lee, 2011) and meta-analysis (Linck et al., 2014) suggested that working memory tasks conducted in L2 may lead to inflated correlations because L2 proficiency was also measured at the same time. Following Indrarathne and Kormos (2017), a composite score of working memory tasks was calculated for each participant based on the component loadings, using the regression method (for correlated factors, Field, 2018) in IBM SPSS Statistics 25. These scores can be used for further statistical analysis, as a way to avoid inflation of statistical significance due to multiple testing on the sample dataset (Tabachnick & Fidell, 2018).

Table 28. Bootstrapped Descriptive and Normality Statistics for Composite Scores of Working Memory Tasks

Mean		0
BCa 95% CI of Mean	*Lower*	-0.24
	Upper	0.24
SD		1
Min		-2.30
Max		2.72
$z_{Skewness}$		0.42
$z_{Kurtosis}$		-0.37

Note. $n = 69$.

Table 28 shows the results of the bootstrapped descriptive statistics for the composite scores based on 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. The *z* scores of skewness and kurtosis were also calculated to assess

data normality (see Figure 7 for the equation). According to the significance level of $p < .01$ with the absolute value of z larger than 2.58 (Tabachnick & Fidell, 2018), the data was normally distributed. As for the interpretation of the composite scores, the larger the score, the higher the working memory capacity. The composite scores can be used in subsequent data analysis.

4.1.4 Post-learning Survey

For the first survey question whether the participants noticed something about the layout of the Chinese characters, pinyin, and English meaning during learning, 10 out of 69 (14.49%) said they did not. As for the preference ratings for the three presentation formats ("Please rate how much you liked each format"), descriptive statistics of mean, standard deviation (SD), minimum (Min), and maximum (Max) were calculated for each presentation format with IBM SPSS Statistics 25. Given the current within-subject design, the Cousineau-Morey method with adjustments (Baguley, 2012) was used to calculate C-M 95% CIs of the mean (see Table 29) and to create C-M 95% CI plots, using R 4.0.2 via RStudio 1.3.1056 (see Figure 15). To assess data normality, z scores of skewness and kurtosis were calculated (see Table 29; see Figure 7 for the equation). Based on the significance level of $p < .01$ with an absolute value of z larger than 2.58 (Tabachnick & Fidell, 2018), the data were normally distributed.

Table 29. Descriptive and Normality Statistics for Preference Ratings

		Horizontal	Vertical	Adjacent
Mean		5.23	4.48	3.22
C-M 95% CI	*Lower*	4.87	4.15	2.80
	Upper	5.60	4.81	3.64
SD		1.77	1.75	2.07
Min		1	1	1
Max		7	7	7
$z_{Skewness}$		-2.40	-0.92	1.69
$z_{Kurtosis}$		-0.79	-1.09	-2.05

Note. $^{*}p < .01$. $n = 69$. The ratings were based on a 7-point Likert scale.

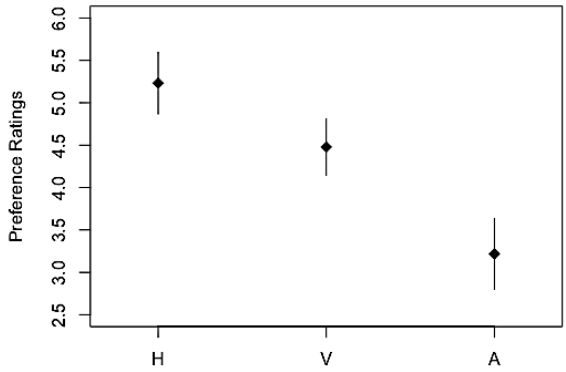

Figure 15. C-M 95% CI plot for preference ratings.

As shown in Table 29 and Figure 15, the preference ratings may be significantly different among the three presentation formats. Specifically, the horizontal format had the highest rating, whereas the adjacent one had the lowest. This indicated that participants liked the horizontal format the most but the adjacent the least.

4.1.5 Chinese Proficiency Test

Bootstrapped descriptive statistics were calculated for the data with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25 (see Table 30). To assess data normality, *z* scores of skewness and kurtosis were calculated (see Figure 7 for the equation). Based on the significance level of $p < .01$ with the absolute value of *z* larger than 2.58 (Tabachnick & Fidell, 2018), the data were not normally distributed.

Table 30. Bootstrapped Descriptive and Normality Statistics for Chinese Proficiency Test

Mean		12.22
BCa 95% CI of Mean	*Lower*	11.35
	Upper	13.03
SD		3.61
Min		1
Max		16
$z_{Skewness}$		-4.45*
$z_{Kurtosis}$		2.11

Note. $^{*}p < .01$. $n = 69$.

4.2 Bivariate Correlations

4.2.1 Fixation Durations and Fixation Counts on Characters, Pinyin, and Meaning

To explore the relationships between characters, pinyin, and meaning in different presentation formats in terms of fixation durations and fixation counts, bootstrapped Pearson's correlations were conducted for the data with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25 (see Table 32). 95% CI plots were also drawn for the correlation coefficient *r* using R 4.0.2 via RStudio 1.3.1056 (see Figure 18).

Table 31. Bootstrapped Pearson's Correlations for Fixation Durations and Fixation Counts Between Characters, Pinyin, and Meaning

				BCa 95% CI of r	
		r	*p*	*Lower*	*Upper*
		Fixation Durations			
H	Character-Meaning (C-M)	-.59*	< .001	-.77	-.37
	Character-Pinyin (C-P)	-.83*	< .001	-.89	-.74
	Pinyin-Meaning (P-M)	.47*	< .001	.21	.70
V	Character-Meaning (C-M)	-.68*	< .001	-.82	-.47
	Character-Pinyin (C-P)	-.85*	< .001	-.93	-.73
	Pinyin-Meaning (P-M)	.59*	< .001	.34	.76
A	Character-Meaning (C-M)	-.69*	< .001	-.85	-.43
	Character-Pinyin (C-P)	-.68*	< .001	-.85	-.41
	Pinyin-Meaning (P-M)	.40*	.001	.07	.65
		Fixation Counts			
H	Character-Meaning (C-M)	-.26*	.030	-.46	-.06
	Character-Pinyin (C-P)	-.44*	< .001	-.61	-.23
	Pinyin-Meaning (P-M)	.54*	< .001	.27	.77
V	Character-Meaning (C-M)	-.28*	.022	-.51	-.01
	Character-Pinyin (C-P)	-.47*	< .001	-.67	-.21
	Pinyin-Meaning (P-M)	.62*	< .001	.39	.79
A	Character-Meaning (C-M)	-.23	.062	-.55	.16
	Character-Pinyin (C-P)	-.24	.053	-.56	.12
	Pinyin-Meaning (P-M)	.46*	< .001	.16	.70

Note. $n = 68$.
*$p < .05$.

As shown in Table 31 and Figure 16, for fixation durations and fixation counts in all

three presentation formats, the characters had significant negative correlations with pinyin and meaning, whereas pinyin and meaning were positively correlated. Notably, the negative correlations of characters to pinyin and meaning did not reach statistical significance for neither fixation durations nor fixation counts in the adjacent format (the whiskers crossing zeros). These results implied a trade-off in attention allocation for fixation durations and fixation counts across three presentation formats: paying more attention to characters would result in less attention to pinyin and meaning. However, attention allocation to pinyin and meaning went in the same direction across the three presentation formats: when participants paid more attention to the pinyin, and they also paid more attention to the meaning.

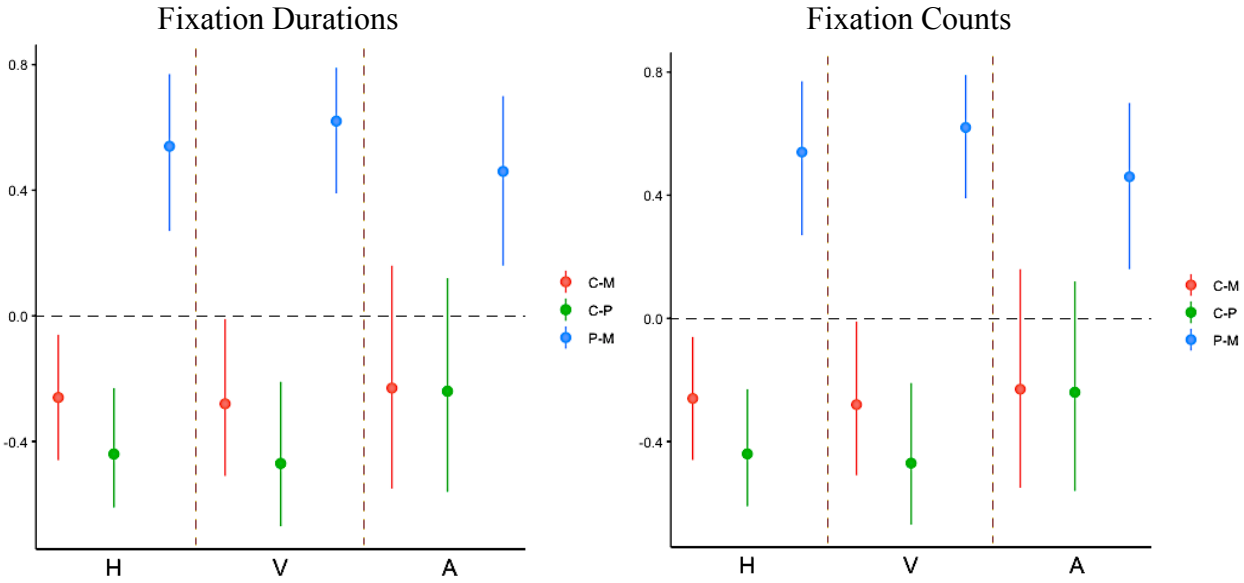

Figure 16. 95% CI plots for Pearson's correlations for fixation durations and fixation counts between characters, pinyin, and meaning. Brown dash-lines were drawn to separate different presentation formats, and a black dash-line was drawn at $r = 0$.

4.2.2 Vocabulary Gain Scores and Fixation Durations/Counts

To explore the relationships between vocabulary gain scores and fixation durations/counts on characters, pinyin, and meaning in different presentation formats,

bootstrapped Pearson's correlations were conducted for the data with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25 (see Table 32). 95% CI plots were also drawn for the correlation coefficient *r* using R 4.0.2 via RStudio 1.3.1056 (see Figure 17).

Table 32. Bootstrapped Pearson's Correlations Between Vocabulary Gain Scores and Fixation Durations/Counts on Characters, Pinyin, and Meaning in Three Presentation Formats

		Fixation Durations			
				BCa 95% CI of r	
		r	*p*	*Lower*	*Upper*
H	Character (C)	.38*	.002	.15	.57
	Pinyin (P)	-.36*	.003	-.56	-.14
	Meaning (M)	-.29*	.015	-.51	-.06
V	Character (C)	.32*	.009	.07	.52
	Pinyin (P)	-.28*	.021	-.50	-.02
	Meaning (M)	-.32*	.007	-.54	-.08
A	Character (C)	.26*	.030	.01	.47
	Pinyin (P)	-.16	.199	-.41	.15
	Meaning (M)	-.25*	.042	-.49	.04
		Fixation Counts			
				BCa 95% CI of r	
		r	*p*	*Lower*	*Upper*
H	Character (C)	.39*	.001	.17	.58
	Pinyin (P)	-.28*	.019	-.50	-.03
	Meaning (M)	-.22	.068	-.45	.04
V	Character (C)	.38*	.001	.10	.60
	Pinyin (P)	-.12	.345	-.36	.15
	Meaning (M)	-.12	.340	-.35	.12
A	Character (C)	.38*	.002	.17	.55
	Pinyin (P)	-.03	.797	-.30	.27
	Meaning (M)	-.17	.155	-.43	.14

Note. $n = 68$.
*$p < .05$.

As shown in Table 32 and Figure 17, within each presentation format, both fixation durations and fixation counts on the characters had significant positive correlations with vocabulary gain scores, whereas the correlations between vocabulary gain scores and fixation

durations/counts on pinyin and meaning may be negative without statistical significance (whiskers crossing zero). This implied that the more time participants looked at the characters, the more vocabulary knowledge they would develop, regardless of the presentation format.

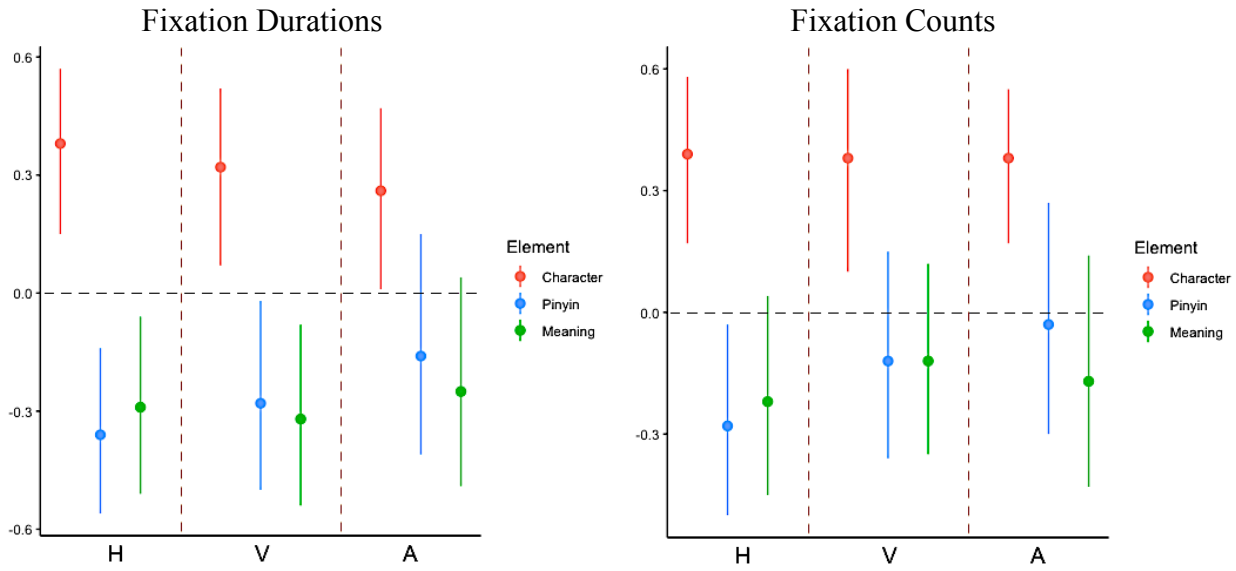

Figure 17. 95% CI plots for Pearson's correlations between vocabulary gain scores and fixation durations/counts on characters, pinyin, and meaning. Brown dash-lines were drawn to separate different presentation formats, and a black dash-line was drawn at $r = 0$.

4.2.3 Vocabulary Gain Scores and L2 Chinese Proficiency

To explore the relationships between vocabulary gain scores and L2 Chinese proficiency, bootstrapped Pearson's correlations were conducted for the data with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. 95% CI plots were also drawn for the correlation coefficient *r* using R 4.0.2 via RStudio 1.3.1056.

As shown in Table 33 and Figure 18, L2 Chinese proficiency had significant positive correlations with vocabulary gain scores in three presentation formats. This indicated that the higher the participants' L2 Chinese proficiency level, the higher vocabulary gain score they would get.

Table 33. Bootstrapped Pearson's Correlations Between Vocabulary Gain Scores and L2 Chinese Proficiency

			BCa 95% CI of r	
	r	*p*	*Lower*	*Upper*
Horizontal	.27*	.027	.04	.47
Vertical	.32*	.007	.08	.53
Adjacent	.41*	< .001	.20	.58

Note. $n = 69$.
*$p < .05$.

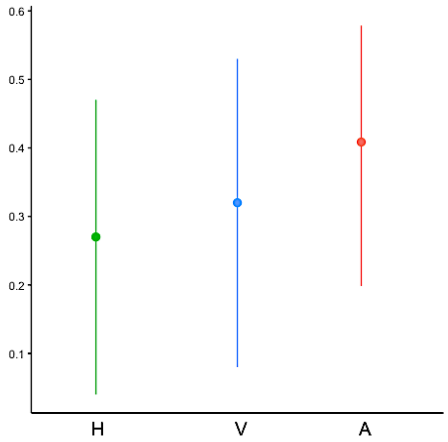

Figure 18. 95% CI plots for Pearson's correlations between vocabulary gain scores and L2 Chinese proficiency.

4.2.4 Vocabulary Gain Scores and Preference Ratings

To explore the relationships between vocabulary gain scores and preference ratings, bootstrapped Pearson's correlations were conducted for the data with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. 95% CI plots were also drawn for the correlation coefficient *r* using R 4.0.2 via RStudio 1.3.1056.

As shown in Table 34 and Figure 19, the vocabulary gain scores did not have significant correlations with preference ratings. This implied that whether participants liked the presentation formats or not may not affect how well they learned with them.

Table 34. Bootstrapped Pearson's Correlations Between Vocabulary Gain Scores and Preference Ratings

			BCa 95% CI of r	
	r	*p*	*Lower*	*Upper*
Horizontal	-.04	.768	-.27	.20
Vertical	.09	.481	-.16	.32
Adjacent	.15	.207	-.07	.37

Note. $n = 69$.
$^{*}p < .05$.

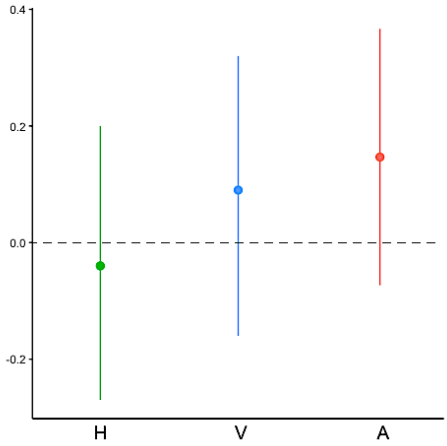

Figure 19. 95% CI plots for Pearson's correlations between vocabulary gain scores and preference ratings. A black dash-line was drawn at $r = 0$.

4.2.5 Fixation Durations/Counts and Preference Ratings

To explore the relationships between fixation durations/counts and preference ratings, bootstrapped Pearson's correlations were conducted for the data with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. 95% CI plots were also drawn for the correlation coefficient *r* using R 4.0.2 via RStudio 1.3.1056.

As shown in Table 35 and Figure 20, the fixation durations/counts did not have significant correlations with preference ratings. This implied that whether participants liked the presentation formats or not may not affect how they paid attention to the three elements.

Table 35. Bootstrapped Pearson's Correlations Between Preference Ratings and Fixation Durations/Counts to Characters, Pinyin, and Meaning in Three Presentation Formats

		Fixation Durations			
				BCa 95% CI of r	
		r	*p*	*Lower*	*Upper*
H	Character	.00	.975	-.26	.25
	Pinyin	-.06	.648	-.33	.22
	Meaning	.14	.259	-.09	.36
V	Character	.11	.367	-.12	.34
	Pinyin	-.08	.542	-.31	.17
	Meaning	-.04	.727	-.27	.19
A	Character	.15	.209	-.10	.37
	Pinyin	-.08	.544	-.29	.18
	Meaning	-.01	.935	-.21	.23
		Fixation Counts			
				BCa 95% CI of r	
		r	*p*	*Lower*	*Upper*
H	Character	.03	.837	-.28	.32
	Pinyin	.04	.769	-.29	.37
	Meaning	.23	.054	.02	.43
V	Character	-.06	.620	-.30	.17
	Pinyin	-.07	.580	-.31	.18
	Meaning	-.01	.924	-.23	.21
A	Character	.12	.316	-.13	.34
	Pinyin	-.05	.689	-.28	.21
	Meaning	-.02	.887	-.26	.25

Note. $n = 68$.
$^{*}p < .05$.

4.2.6 Vocabulary Gain Scores and Working Memory Capacities

To explore the relationships between vocabulary gain scores and working memory capacities, bootstrapped Pearson's correlations were conducted for the data (imputed square-rooted data as well as the composite scores of four working memory tasks, see 4.1.3 Working Memory Tasks) with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25 (see Table 36). Notably, regarding the data of number letter and Stroop tasks, all RT differences were reversed (reversion was performed when calculating the

composite scores, see 4.1.3 Working Memory Tasks), so that for all working memory measures, the larger the value, the better the task performance, following Friedman et al. (2008). 95% CI plots were also drawn for the correlation coefficient *r* using R 4.0.2 via RStudio 1.3.1056 (see Figure 21).

As shown in Table 36, for all three presentation formats, vocabulary gain scores had significant positive correlations with the composite scores as well as two task performance: number letter and Stroop for the horizontal and the adjacent formats, and letter memory and Stroop for the vertical format. However, when checking the 95% CI plot in Figure 21, for the horizontal and the vertical formats, only the digit span (DS) task performance did not have a significant correlation with the vocabulary gain scores (whiskers crossing zero). For the adjacent format, the performance of either digit span (DS) or letter memory (LM) task did not correlate significantly with vocabulary gain scores. The positive correlations indicated the better the working memory capacities, the higher vocabulary gains, regardless of presentation formats.

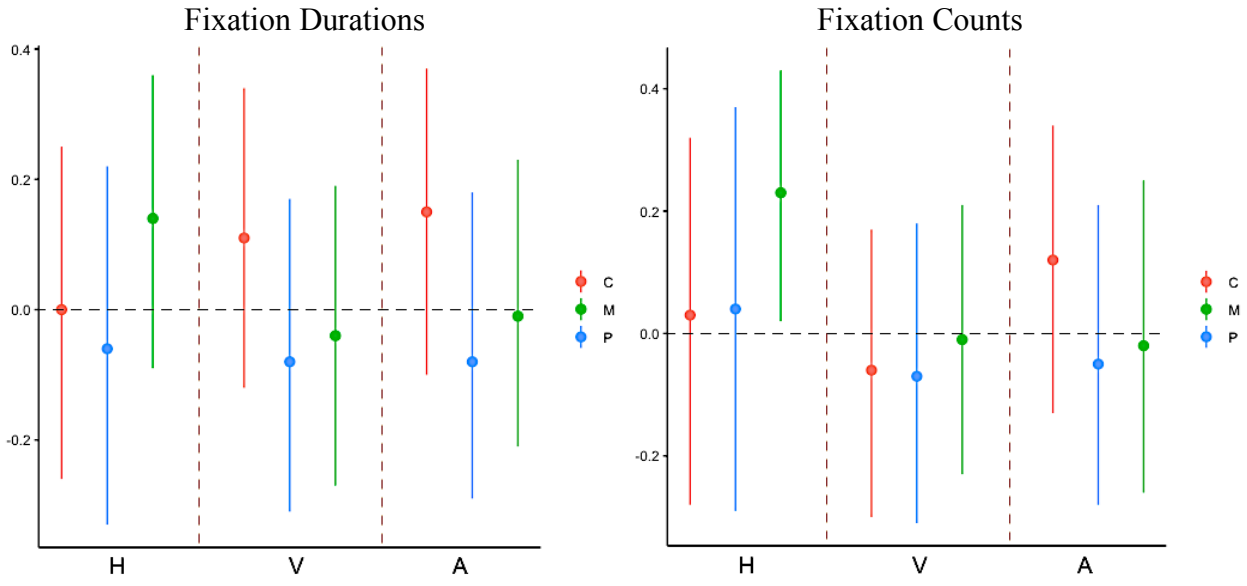

Figure 20. 95% CI plots for Pearson's correlations between preference ratings and fixation durations/counts to characters, pinyin, and meaning. Brown dash-lines were drawn to separate different presentation formats, and a black dash-line was drawn at $r = 0$.

Table 36. Bootstrapped Pearson's Correlations Between Vocabulary Gain Scores and Imputed Square-rooted Data and Composite Scores of Four Working Memory Tasks

				BCa 95% CI of r	
		r	*p*	*Lower*	*Upper*
H	Digit Span (DS)	.18	.138	-.05	.40
	Letter Memory (LM)	.23	.062	.03	.40
	Number Letter (NL)	.25*	.040	.02	.47
	Stroop (ST)	.39*	.001	.20	.56
	Composite Score (CS)	.38*	.001	.20	.54
V	Digit Span (DS)	.13	.277	-.10	.36
	Letter Memory (LM)	.29*	.018	.06	.48
	Number Letter (NL)	.24	.052	.01	.44
	Stroop (ST)	.39*	.001	.19	.57
	Composite Score (CS)	.38*	.001	.18	.54
A	Digit Span (DS)	.16	.187	-.07	.38
	Letter Memory (LM)	.17	.158	-.05	.38
	Number Letter (NL)	.28*	.020	.03	.51
	Stroop (ST)	.30*	.013	.06	.52
	Composite Score (CS)	.33*	.006	.12	.50

Note. $n = 68$.
*$p < .05$.

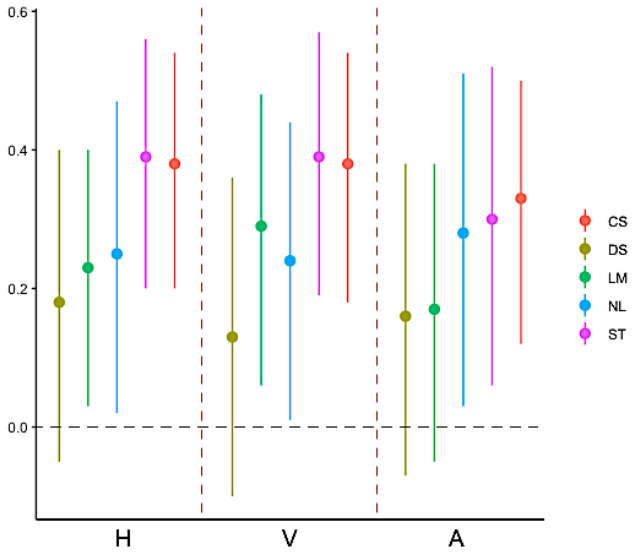

Figure 21. 95% CI plots for Pearson's correlations between vocabulary gain scores and imputed square-rooted data and composite scores of four working memory tasks. Brown dash-lines were drawn to separate different presentation formats, and a black dash-line was drawn at $r = 0$.

4.2.7 Fixation Durations/Counts and Working Memory Capacities

To explore the relationships between fixation durations/counts and working memory capacities, bootstrapped Pearson's correlations were conducted for the data (imputed square-rooted data as well as the composite scores of four working memory tasks, see 4.1.3 Working Memory Tasks) with 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016), using IBM SPSS Statistics 25. Notably, regarding the data of number letter and Stroop tasks, all RT differences were reversed (reversion was performed when calculating the composite scores, see 4.1.3 Working Memory Tasks), so that for all working memory measures, the larger the value, the better the task performance, following Friedman et al. (2008). 95% CI plots were also drawn for the correlation coefficient *r* using R 4.0.2 via RStudio 1.3.1056. Table 37 and Figure 22 display the results and visuals for fixation durations, and Table 38 and Figure 23 display those for fixation counts.

Table 37. Bootstrapped Pearson's Correlations Between Fixation Durations and Imputed Square-rooted Data and Composite Scores of Four Working Memory Tasks

					BCa 95% CI of r	
			r	*p*	*Lower*	*Upper*
H	Character (C)	Digit Span (DS)	.12	.315	-.09	.32
		Letter Memory (LM)	.08	.503	-.17	.34
		Number Letter (NL)	.21	.087	-.05	.45
		Stroop (ST)	.19	.123	-.04	.41
		Composite Score (CS)	.22	.079	-.04	.45
	Pinyin (P)	Digit Span (DS)	-.19	.123	-.41	.07
		Letter Memory (LM)	-.21	.082	-.45	.05
		Number Letter (NL)	-.24*	.047	-.46	.01
		Stroop (ST)	-.22	.071	-.39	-.03
		Composite Score (CS)	-.32*	.009	-.53	-.05
	Meaning (M)	Digit Span (DS)	.07	.564	-.14	.26
		Letter Memory (LM)	-.07	.554	-.29	.14
		Number Letter (NL)	-.10	.415	-.31	.12
		Stroop (ST)	-.02	.861	-.25	.19
		Composite Score (CS)	-.04	.721	-.26	.17

Table 37 (cont'd)

					BCa 95% CI of r	
			r	*p*	*Lower*	*Upper*
V	Character (C)	Digit Span (DS)	.18	.140	-.05	.41
		Letter Memory (LM)	.06	.610	-.16	.30
		Number Letter (NL)	.16	.197	-.08	.36
		Stroop (ST)	.09	.480	-.11	.29
		Composite Score (CS)	.18	.151	-.06	.41
	Pinyin (P)	Digit Span (DS)	-.16	.194	-.37	.06
		Letter Memory (LM)	-.13	.307	-.35	.09
		Number Letter (NL)	-.15	.209	-.34	.06
		Stroop (ST)	-.07	.577	-.25	.13
		Composite Score (CS)	-.19	.127	-.39	.03
	Meaning (M)	Digit Span (DS)	-.18	.153	-.41	.07
		Letter Memory (LM)	-.16	.208	-.35	.03
		Number Letter (NL)	-.14	.256	-.34	.08
		Stroop (ST)	-.08	.509	-.27	.11
		Composite Score (CS)	-.21	.094	-.41	.00
					BCa 95% CI of r	
			r	*p*	*Lower*	*Upper*
A	Character (C)	Digit Span (DS)	.13	.290	-.09	.36
		Letter Memory (LM)	.15	.239	-.11	.43
		Number Letter (NL)	.20	.097	-.06	.41
		Stroop (ST)	.12	.350	-.08	.31
		Composite Score (CS)	.22	.077	-.01	.46
	Pinyin (P)	Digit Span (DS)	-.17	.169	-.39	.06
		Letter Memory (LM)	-.29*	.016	-.53	-.04
		Number Letter (NL)	-.19	.124	-.37	.03
		Stroop (ST)	-.03	.801	-.24	.20
		Composite Score (CS)	-.26*	.033	-.50	-.01
	Meaning (M)	Digit Span (DS)	-.03	.836	-.25	.19
		Letter Memory (LM)	-.05	.697	-.27	.14
		Number Letter (NL)	-.14	.269	-.34	.09
		Stroop (ST)	-.14	.253	-.34	.06
		Composite Score (CS)	-.12	.319	-.33	.09

Note. $n = 68$.
*$p < .05$.

As shown in Table 37, for the horizontal format, fixation durations on pinyin had significant negative correlations with the number letter task performance and the composite scores. However, when checking the 95% CI plots in Figure 22, for the horizontal format,

fixation durations on pinyin may have significant negative correlations with Stroop (ST) task performance and the composite scores instead. For the adjacent format, both Table 37 and Figure 22 confirm that fixation durations on pinyin had significant negative correlations with the letter memory task performance and the composite scores. These results indicated that working memory capacities may affect allocating attention to pinyin in the horizontal and adjacent formats. Specifically, the higher working memory capacities, the less attention to pinyin.

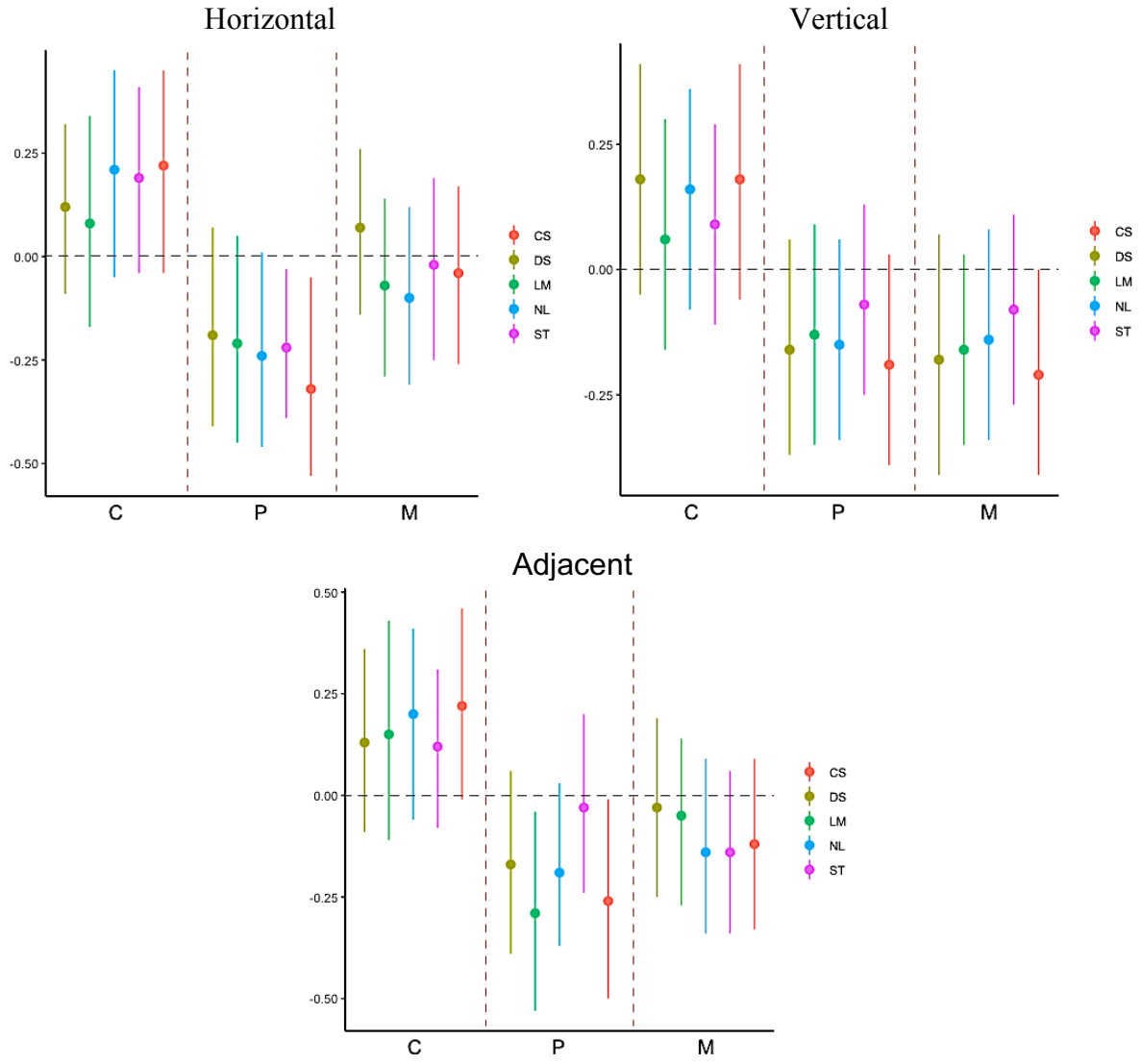

Figure 22. 95% CI plots for Pearson's correlations between fixation durations and imputed square-rooted data and composite scores of four working memory tasks. Brown dash-lines were drawn to separate different presentation formats, and a black dash-line was drawn at $r = 0$.

Table 38. Bootstrapped Pearson's Correlations Between Fixation Counts and Imputed Square-rooted Data and Composite Scores of Four Working Memory Tasks

					BCa 95% CI of r	
			r	*p*	*Lower*	*Upper*
H	Character (C)	Digit Span (DS)	.21	.093	.00	.39
		Letter Memory (LM)	.20	.106	-.02	.40
		Number Letter (NL)	.17	.155	-.08	.42
		Stroop (ST)	.23	.056	.02	.45
		Composite Score (CS)	.30*	.013	.10	.48
	Pinyin (P)	Digit Span (DS)	-.10	.435	-.34	.17
		Letter Memory (LM)	-.09	.485	-.33	.17
		Number Letter (NL)	-.13	.291	-.35	.12
		Stroop (ST)	-.13	.287	-.31	.07
		Composite Score (CS)	-.16	.192	-.40	.11
	Meaning (M)	Digit Span (DS)	.06	.602	-.19	.30
		Letter Memory (LM)	.00	.978	-.23	.22
		Number Letter (NL)	-.05	.668	-.29	.19
		Stroop (ST)	.04	.756	-.17	.25
		Composite Score (CS)	.02	.858	-.24	.28
					BCa 95% CI of r	
			r	*p*	*Lower*	*Upper*
V	Character (C)	Digit Span (DS)	.31*	.011	.08	.50
		Letter Memory (LM)	.20	.097	-.02	.43
		Number Letter (NL)	.17	.179	-.06	.37
		Stroop (ST)	.12	.346	-.12	.34
		Composite Score (CS)	.30*	.015	.10	.48
	Pinyin (P)	Digit Span (DS)	-.06	.624	-.30	.18
		Letter Memory (LM)	-.04	.762	-.28	.20
		Number Letter (NL)	-.15	.220	-.38	.11
		Stroop (ST)	-.03	.840	-.22	.18
		Composite Score (CS)	-.10	.439	-.33	.16
	Meaning (M)	Digit Span (DS)	-.06	.654	-.30	.19
		Letter Memory (LM)	-.03	.810	-.24	.18
		Number Letter (NL)	-.19	.131	-.43	.08
		Stroop (ST)	.05	.699	-.14	.23
		Composite Score (CS)	-.08	.539	-.31	.16
					BCa 95% CI of r	
			r	*p*	*Lower*	*Upper*
A	Character (C)	Digit Span (DS)	.26*	.032	.04	.47
		Letter Memory (LM)	.28*	.022	.03	.52
		Number Letter (NL)	.14	.261	-.09	.34
		Stroop (ST)	.12	.341	-.10	.33
		Composite Score (CS)	.30*	.013	.10	.49

Table 38 (cont'd)

Pinyin (P)	Digit Span (DS)	-.09	.462	-.32	.13
	Letter Memory (LM)	-.12	.329	-.39	.14
	Number Letter (NL)	-.19	.120	-.43	.09
	Stroop (ST)	.02	.848	-.18	.24
	Composite Score (CS)	-.14	.262	-.38	.13
Meaning (M)	Digit Span (DS)	-.01	.947	-.23	.24
	Letter Memory (LM)	.02	.900	-.22	.22
	Number Letter (NL)	-.13	.310	-.35	.12
	Stroop (ST)	-.06	.639	-.26	.15
	Composite Score (CS)	-.06	.646	-.29	.18

Note. $n = 68$.
$^{*}p < .05$.

Table 38 and Figure 23 display the results and visuals for fixation counts. For the horizontal format, Table 38 shows that the composite score had a significant positive correlation with fixation counts on characters, whereas in Figure 23, in addition to the composite scores, Stroop task performance also had a significant positive correlation with fixation counts to characters. For the vertical format, both Table 38 and Figure 23 support that fixation counts on characters had significant positive correlations with the composite scores and digit spans. For the adjacent format, both Table 38 and Figure 23 suggest that fixation counts on characters had significant positive correlations with composite scores, digit spans, as well as letter memory task performance. These results indicated that the higher working memory capacities, the more times the participants would look at the characters, regardless of presentation formats.

4.3 Summary of Descriptive Statistics and Bivariate Correlations

To explore the potential relationships among presentation formats, working memory capacities, learner attention, and learning outcomes, descriptive and normality statistics as well as bivariate correlations were calculated with the quantitative data indices. Results suggested that

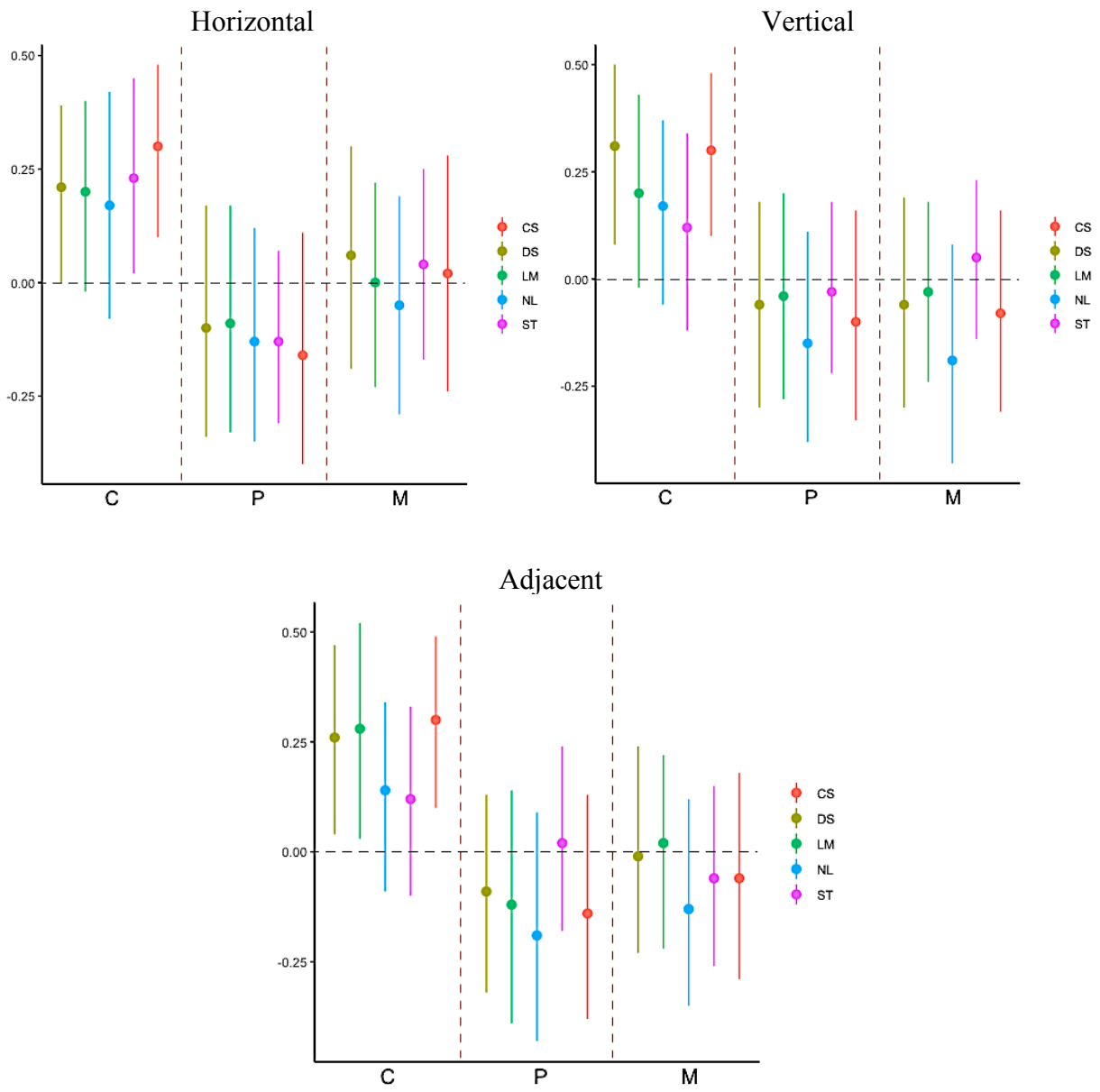

Figure 23. 95% CI plots for Pearson's correlations between fixation counts and imputed square-rooted data and composite scores of four working memory tasks. Brown dash-lines were drawn to separate different presentation formats, and a black dash-line was drawn at $r = 0$.

vocabulary gain scores may be affected by presentation formats (see 4.1.1), vocabulary test formats (see 4.1.1), learner attention (see 4.2.2), L2 Chinese proficiency (see 4.2.3), and working memory capacities (see 4.2.6). Although learner preference may differ among the presentation formats (see 4.1.4), it may not affect vocabulary gain scores (see 4.2.4) or learner attention (see 4.2.5). Regarding learner attention, presentation formats may affect how learners pay attention to

the characters, pinyin, and meaning (see 4.1.2 & 4.2.1), and working memory capacities may affect how many times they look at the characters (see 4.2.7).

Building on the results of descriptive statistics and bivariate correlations, mixed effects modeling was conducted to unmask the effects of presentation formats and working memory capacities by accounting for their relationships with other variables at the same time (Cunnings, 2012; Linck & Cunnings, 2015), in addition to repeated-measures ANOVA, so as to unveil the full picture of these variables and their relationships.

4.4 Overview of Mixed Effects Models for RQs

Although there are eight RQs in total, based on the outcome variable, the mixed effects models for the RQs can be divided into two categories, with vocabulary gain scores (for RQs 1, 3, 5, & 7) and attention data of fixation durations and fixation counts (for RQs 2, 6, & 8) as the outcome variable respectively. Specifically, the final models of RQs 1 and 2 provided the baseline models for subsequent modeling building for other related RQs. All mixed effects models were built with R 4.0.2 via RStudio 1.3.1056. For RQ 4 on preference ratings, repeated-measures ANOVA was conducted to compare the three presentation formats.

4.5 RQ Set A Focusing on Presentation Formats

4.5.1 RQ 1. What is the relationship between *presentation formats* (i.e., horizontal, vertical, and adjacent) and *learning outcomes* (as assessed by a bilingual vocabulary test) in L2 Chinese vocabulary learning, taking L2 Chinese proficiency and test formats into consideration?

Distribution of the Outcome Variable. Basic steps in statistical modeling include

starting with deciding a distribution for the outcome variable, before feeding the predictors into the model (Cunnings & Finlayson, 2015; Gelman & Hill, 2007). As mentioned in 3.5.2 Overall Analytical Approach and Statistical Methods, one advantage of mixed effects modeling is the use of raw data without aggregation (Cunnings, 2012; Cunnings & Finlayson, 2015; Linck & Cunnings, 2015), and for the current RQ, vocabulary gain scores at the item level provided data for the outcome variable in the mixed effects model, but further consideration is needed on their distribution. The decision on which distribution to choose should be based on the available knowledge on the outcome variable in the first place (Zuur et al., 2009).

As described in 3.3.2 Pretest and Posttest (see Tables 12 & 13), 0/1 scoring was adopted for four recognition tasks (C2M_rcg, M2C_rcg, C2P_rcg, and M2P_rcg) and one recall task (C2M_rcl), whereas fraction scoring between 0 and 1 (e.g., 0.5) was applied to three recall tasks (M2C_rcl, C2P_rcl, M2P_rcl). For the five tasks using 0/1 scoring, binomial distribution is generally appropriate to describe the dichotomous possibilities (e.g., Yes/No) (Cunnings & Finlayson, 2015; Gelman & Hill, 2007; Zuur et al., 2009). For the three tasks using fraction scoring between 0 and 1, a histogram was drawn for the item-level vocabulary gain scores with R 4.0.2 via RStudio 1.3.1056 (see Figure 24), so as to provide initial visualization of data distribution (Larson-Hall, 2016). As in Figure 24, excessive zeros (96.94%) piled up on the left part of the histogram, with comparatively small numbers of fractions on the right part. To further examine the distribution of the non-zero data, another histogram was drawn using R 4.0.2 via RStudio 1.3.1056 (see Figure 25). As in Figure 25, the non-zero data were all positive, mostly continuous, and left skewed, which can be categorized as *semicontinuous data*.

Semicontinuous data refer to datasets that contain a large number of zeros and a continuous distribution of positive values, often with right skewness (Farewell, Long, Tom, Yiu,

& Su, 2017; Liu, Shih, Strawderman, Zhang, Johnson, & Chai, 2019; Neelon & O'Malley, 2019; Neelon, O'Malley, & Smith, 2016a) and sometimes with left skewness (e.g., Elsabry & Sumikura, 2020). *Two-part models* have received much attention in recent years as an effective tool to model semicontinuous data, and several review articles and tutorials have provided guidance on building two-part mixed effects models for semicontinuous data in biomedical, economic and ecological research (see Farewell et al., 2017; Liu et al., 2019; Neelon et al., 2016a, 2016b). The two-part models, as indicated by its name, identify the data points as zero responses and non-zero responses, and use two separate sub-models for each type of responses (Farewell et al., 2017; Liu et al., 2019; Neelon & O'Malley, 2019; Neelon et al., 2016a). Specifically, one sub-model is to account for the occurrence of zeros (i.e., binary part), and the other one is for the non-zero values (i.e., continuous part) (Farewell et al., 2017; Neelon & O'Malley, 2019; Neelon et al., 2016a). Bernoulli distribution, which is a specific form of binomial distribution (N = 1) (Zuur et al., 2009), is usually used for the binary part of the two-part model (Neelon & O'Malley, 2019; Neelon et al., 2016a; Zuur & Ieno, 2016). For the continuous part, lognormal distribution or generalized gamma distribution can be used to model the non-zero data (Neelon & O'Malley, 2019; Neelon et al., 2016a; Zuur & Ieno, 2016), with the latter being more flexible in dealing with skewness and heteroscedasticity (Liu et al., 2019).

Given the semicontinuous nature of the item-level gain scores of the three recall tasks (M2C_rcl, C2P_rcl, M2P_rcl), a two-part mixed effects model was adopted, with Bernoulli distribution for the binary part (Zuur & Ieno, 2016). To decide between lognormal distribution and generalized gamma distribution for the continuous part, a histogram was drawn for the natural-log-transformed (ln) gain scores using R 4.0.2 via RStudio 1.3.1056 (see Figure 25). Compared with Figure 24, the histogram in Figure 25 was more left-skewed, which indicated a

generalized gamma distribution may be more appropriate.

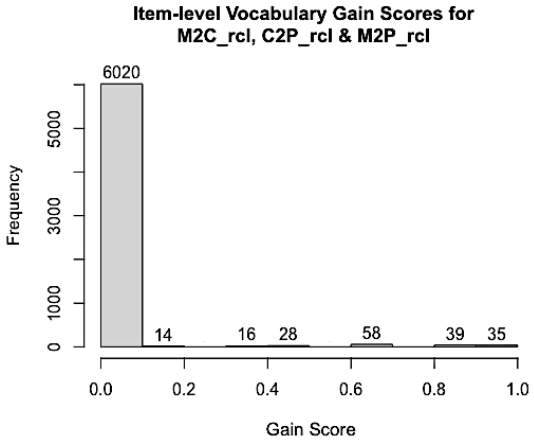

Figure 24. Histogram of item-level vocabulary gain scores for M2C_rcl, C2P_rcl, and M2P_rcl.

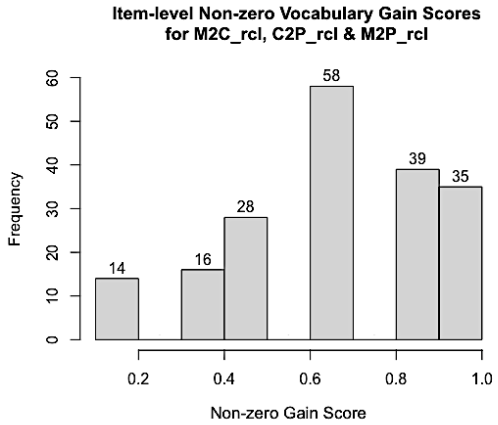

Figure 25. Histogram of item-level non-zero vocabulary gain scores for M2C_rcl, C2P_rcl, and M2P_rcl.

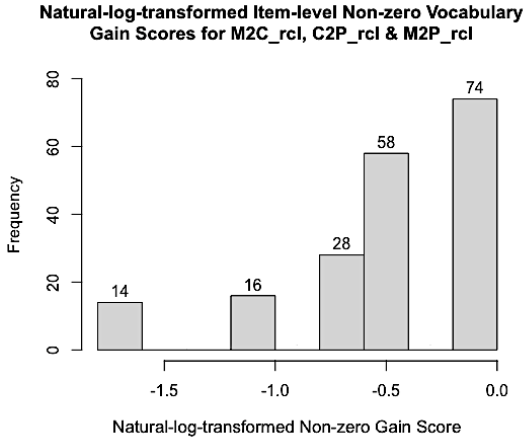

Figure 26. Histogram of natural-log-transformed item-level non-zero vocabulary gain scores for M2C_rcl, C2P_rcl, and M2P_rcl.

Therefore, based on the scoring methods of the vocabulary test items, for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl with 0/1 scoring, binomial distribution was chosen to build the *mixed logit model*, which is an extension of logistic regression to also account for random effects (Jaeger, 2008). For M2C_rcl, C2P_rcl, and M2P_rcl with fraction scoring between 0 and 1, a two-part mixed effects model was built with Bernoulli distribution for the binary part and generalized gamma distribution for the continuous part (see Zuur & Ieno, 2016).

Fixed Effects. After deciding the distribution of the outcome variable and the model type, appropriate predictors need to be selected for inclusion into the model (Cunnings & Finlayson, 2015; Gelman & Hill, 2007). As indicated in the current RQ, presentation format was a predictor of major interest, with L2 Chinese proficiency level and vocabulary test format as additional predictors. The inclusion of these two additional predictors was also supported by the correlations between vocabulary gain scores and L2 Chinese proficiency level (see Table 33 in 4.2.3 Vocabulary Gain Scores and L2 Chinese Proficiency), as well as the results of C-M 95% CI plot for the vocabulary gain scores of the eight test formats (see Figure 13 in 4.1.1 Pretest and Posttest). During model building for the current RQ, the fixed effects were assumed to come from presentation formats, L2 Chinese proficiency, and vocabulary test formats.

Random Effects. The random effects were assumed to come from the participants recruited and the words selected for this study. Regarding specifying the random effects structure in mixed effects models, L2 researchers (Cunnings & Finlayson, 2015; Link & Cunnings, 2015) followed Barr, Levy, Scheepers, and Tily (2013) as well as Barr (2013) in recommending maximal models to include all random effects structures justified by the experimental design for confirmatory research (i.e., hypothesis testing). Specifically, in terms of the random effects for control predictors, based on limited available information about this issue, Barr et al. (2013)

suggested that "it is not essential for one to specify random effects for control predictors to avoid anticonservative inference, as long as interactions between the control predictors and the factors of interest are not present in the model (or justified by the data)" (p. 275). Another issue is random effects for interactions, and by updating the guidelines in Barr et al. (2013), Barr (2013) advised that "models testing interactions in designs with replications should include random slopes for the highest-order combination of within-unit factors subsumed by each interaction" (p. 1). That is, for repeated-measures design, when two or more within-subject and/or within-item factors form interactions, a random slope should be assigned to the-highest-order interaction term. I followed these general guidelines in building mixed effects models for the RQs.

Data Preparation. To prepare the data for model building, presentation formats and vocabulary test formats as categorical predictors were recorded using deviation coding, which is recommended by Barr et al. (2013) as a preferred coding scheme to assess main effects in mixed effects models. Specifically, I followed the guidance by UCLA's Institute for Digital Research & Education Statistical Consulting (https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/) as well as Sonderegger, Wagner, and Torreira (2018) in using the R function contr.sum to recode the data. In order to avoid collinearity among predictors and to increase interpretability of the results, I followed Gelman and Hill (2007) in calculating standardized *z* scores for the L2 Chinese proficiency scores using the following equation:

$$z\ score = \frac{Observed\ Value - Mean}{2 * Standard\ Deviation}$$

Figure 27. Equation for calculating standardized z scores for continuous predictors. Adapted from Gelman and Hill (2007, p. 54).

Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl. I used the glmer function from the lme4 package (version 1.1-23) (Bates, Mächler, Bolker, &

Walker, 2015) to build the mixed logit model, by specifying the binomial family, using the "bobyqa" optimizer (see Link & Cunnings, 2015), and setting the maximum iterations as 200,000 (see Miller, 2018). During model building, I started with the maximal model for both the fixed effects (i.e., include all main effects and interactions of theoretical interest in this study) and the random effects (Barr et al., 2013). Regarding the situation where the model failed to converged, Cunnings and Finlayson (2015) explained the non-convergence is usually caused by complex random effects structures and suggested simplifying them to achieve model convergence. Therefore, when non-convergence happened, I simplified the random effects structure by following the general guidelines by Barr et al. (2013). I also followed Cunnings and Finlayson (2015) in locating the lowest variance estimate of the random effects in the non-converged model and refitting the model by removing that random effect. When the maximal feasible random effects structure was attained (i.e., the model converged), I checked the *t* statistics and *p* values of the fixed effects, and then removed one of the non-significant fixed effects to build a more parsimonious model. Then I used the anova function to compare the models with different fixed effects structures based on AIC (Akaike Information Criterion) (Cunnings & Finlayson, 2015) and selected a final model, that is, the most parsimonious model.

The formula for the final mixed logit model was:

Gain ~ PF + Prof + TF + (1 + PF | ID) + (1 + PF | Word)

In the formula, the outcome variable was item-level vocabulary gain scores (Gain; 0 or 1), and the fixed effects were presentation formats (PF; 3 levels), L2 Chinese proficiency (Prof; standardized), and vocabulary test formats (TF; 5 levels). The random effects structure consisted of a random intercept and a random slope of presentation format varied by participant (ID) and by word (Word) respectively.

Table 39. Results of Mixed Logit Model for RQ 1

Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.588*	0.169	-3.483	< .001
PF-A		0.109	0.056	1.953	.051
PF-H		-0.074	0.057	-1.283	.200
Prof		0.817*	0.245	3.331	< .001
M2C_rcg		0.833*	0.047	17.883	< .001
M2P_rcg		-0.250*	0.047	-5.305	< .001
C2M_rcl		-0.978*	0.052	-18.920	< .001
C2M_rcg		0.986*	0.047	20.929	< .001
Random Effects					
		Variance	*Std.Dev.*		
ID	(Intercept)	0.970	0.985		
	PF-A	0.077	0.277		
	PF-H	0.074	0.272		
Word	(Intercept)	0.415	0.644		
	PF-A	0.025	0.157		
	PF-H	0.030	0.172		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = L2 Chinese Proficiency.

Table 39 shows the results of the final model. As mentioned previously, deviation coding was used for the categorical variables: presentation formats and vocabulary test formats. According to UCLA's webpage (see Data Preparation in this RQ for the link), with deviation coding, the mean of each level of the outcome variable is compared to its grant mean, which is calculated by adding the means of all levels and then dividing the sum by the total number of levels. In addition, the estimate for a level of the outcome variable is calculated by using the mean of this level to minus the grand mean. Following these guidelines, the mean vocabulary gain scores (log-odds) of each presentation format were calculated for the current results: Adjacent = (-0.588) + (0.109) = -0.479, Horizontal = (-0.588) + (-0.074) = -0.662, and Vertical = (-0.588)*3 – (-0.479) – (-0.662) = -0.623. These results indicated that the mean of the adjacent format (-0.479) was almost higher than the grand mean (-0.588) of all presentation formats with

marginal significance (p = .051), when L2 Chinese proficiency and vocabulary test formats were controlled. When presentation formats and vocabulary test formats were taken into consideration, the higher L2 Chinese proficiency level, the higher vocabulary gains (Estimate = 0.817, p < .001). A main effect of vocabulary test formats was also found with statistical significance. These findings were similar to the results of the descriptive statistics (see 4.1.1 for presentation formats and vocabulary test formats; 4.2.3 for L2 Chinese proficiency).

Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl. Following Zuur and Ieno's (2016) guidelines and examples in building zero-inflated mixed effects models, I used the glmer function from the lme4 package (version 1.1-23) (Bates et al., 2015) to model the continuous part (i.e., non-zero values) of the two-part mixed effects model with generalized gamma distribution (a log link to ensure positive values), and to model the binary part (i.e., zero values) with Bernoulli distribution, by using the "bobyqa" optimizer (see Link & Cunnings, 2015) and setting the maximum iterations as 200,000 (see Miller, 2018). The steps in model building and selection were the same as those for the mixed logit model.

The formulae for the final two-part mixed effects model were:

Continuous: Gain ~ 1 + (1 | ID) + (1 | Word)

Binary: Gain ~ Prof + TF + (1 | ID) + (1 | Word)

For both parts, the outcome variable was item-level vocabulary gain scores (Gain; fraction score between 0 and 1). Regarding the continuous part, no fixed effects were found to be significant, and the random effects structure consisted of a random intercept varied by participant (ID) and by word (Word) respectively. In terms of the binary part, the fixed effects were L2 Chinese proficiency (Prof; standardized) and vocabulary test formats (TF; 3 levels). The random effects structure for the binary part consisted of a random intercept varied by participant (ID) and

by word (Word) respectively.

Table 40 shows the results of the final model. For the continuous part, only the intercept was found as significant in the fixed effects component. This indicated that when there were vocabulary gains, the extent of gains (i.e., exact gain scores) were not associated with presentation formats, L2 Chinese proficiency, or vocabulary test formats. As for the binary part, when considering whether vocabulary gains would be obtained or not, L2 Chinese proficiency was found as a significant predictor (Estimate = 0.785, $p < .001$), indicating the higher Chinese proficiency, the more likely (log-odds) that there would be vocabulary gains. Vocabulary test formats were also found as a significant predictor, implying that different formats may pose different levels of difficulty.

To summarize, for the vocabulary test formats of C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl (0/1 scoring), the adjacent format was found to facilitate vocabulary learning, and learning outcome was also associated with L2 Chinese proficiency and vocabulary test formats. For the vocabulary test formats of M2C_rcl, C2P_rcl, and M2P_rcl (fraction scoring between 0 and 1), whether there would be learning gains or not was associated with L2 Chinese proficiency and vocabulary test formats, but not presentation formats. However, when learning gains were obtained, the extent of gains was not associated with presentation formats, L2 Chinese proficiency, or vocabulary test formats. Considering all vocabulary test formats together, L2 Chinese proficiency and vocabulary test formats may consistently affect the learning outcome.

Table 40. Results of Two-Part Mixed Effects Model for RQ 1

		Continuous				Binary			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.581*	0.086	-6.781	< .001	-5.378*	0.318	-16.911	< .001
Prof						0.785*	0.382	2.055	.040
M2C_rcl						-1.384*	0.242	-5.709	< .001
M2P_rcl						-0.541*	0.195	-2.782	.005
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.029	0.171			1.655	1.286		
Word	Intercept	0.020	0.142			0.918	0.958		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Prof = L2 Chinese Proficiency.

4.5.2 RQ 2. What is the relationship between *presentation formats* (i.e., horizontal, vertical, and adjacent) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in L2 Chinese vocabulary learning?

Distribution of the Outcome Variable. As mentioned previously, the first step in building mixed effects models is to decide a data distribution for the outcome variable (Cunnings & Finlayson, 2015; Gelman & Hill, 2007). Therefore, I used R 4.0.2 via RStudio 1.3.1056 to create a histogram for the data of fixation durations and fixation counts respectively (see Figure 28), so as to provide initial visualization of the data (Larson-Hall, 2016). Specifically, the data of fixation durations and fixation counts were at element level (i.e., for the characters, pinyin, or English meaning of a Chinese word), with totally three data points for one Chinese word during each of the two times of learning (see 3.4 Procedure). As shown in Figure 28, the data of fixation durations and fixation counts were right skewed and did not follow normal distributions. Additionally, to detect potential issues of excessive zeros in the data, I also checked the number of zero values in the two datasets: 44 (0.36%) for fixation durations and fixation counts respectively.

Regarding fixation durations, psychology literature has recommended log transformation to achieve near-normal distribution, before feeding the data into mixed effects models with fixation durations as the outcome variable (e.g., Hohenstein, Matuschek, & Klieg, 2017; Nuthmann, 2017). Therefore, I created another histogram for non-zero values that were natural-log-transformed (see Figure 29). However, the data were left skewed and still far from normally distributed. As Zuur et al. (2009) and Zuur and Ieno (2016) suggested, generalized gamma distribution provides an alternative for positive continuous data that do not follow normal distributions. Since generalized gamma distribution does not allow zeros (Zuur et al., 2009; Zuur

& Ieno, 2016), I converted zero values of fixation durations into missing values before fitting the model with a generalized gamma distribution (a log link to ensure positive values). As mentioned previously, only 44 zeros (0.36%) existed in the data, and the loss of these data may not have a large impact on the overall results.

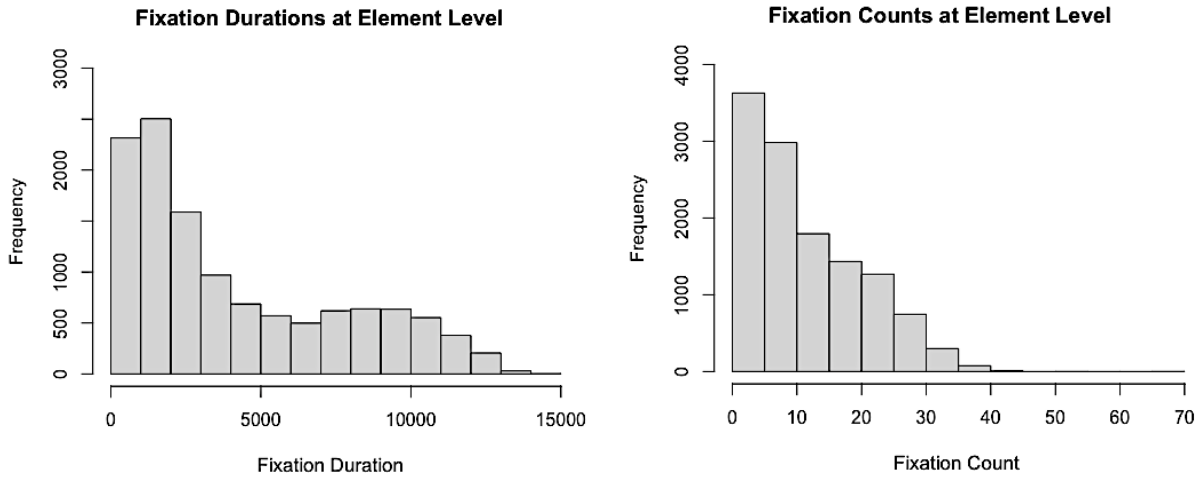

Figure 28. Histograms of fixation durations and fixation counts at element level, and transitions at unidirectional level.

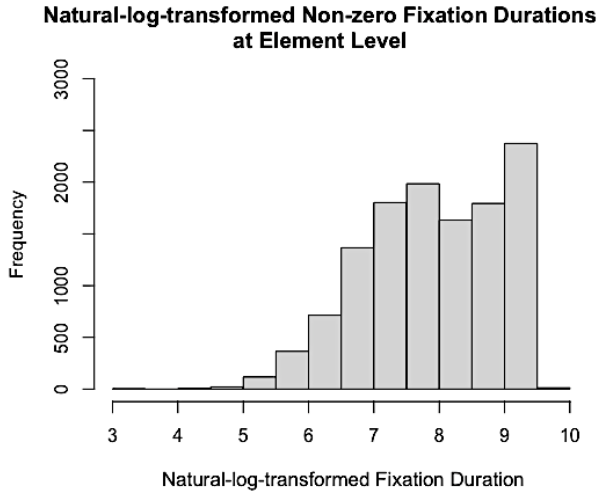

Figure 29. Histogram of natural-log-transformed fixation durations at element level.

In terms of fixation counts, recent eye-tracking literature has adopted negative binomial distribution for mixed effects models with fixation counts as the outcome variable (e.g., Hunt,

Stuart, Nell, Hausdorff, Galna, Rochester, & Alcock, 2018; Man & Harring, 2020; Noland Weiner, Gao, Cook, & Nelessen, 2017). Compared with the often-used Poisson distribution for count data, negative binomial distribution is recommended for its further flexibility in allowing under/over-dispersion, that is, the variance is unequal to the mean (Man & Harring, 2019). Specifically, when the variance is greater than the mean, overdispersion occurs, whereas underdispersion refers to when the variance is lower than the mean (Hardin & Hilbe, 2018). Usually overdispersion is more common than underdispersion (Zurr et al., 2009; Zurr & Ieno, 2016). A dispersion parameter can be calculated for a fitted Poisson model, as a way to check whether under/over-dispersion exists (Gelman & Hill, 2007; Zuur et al., 2009; Zuur & Ieno, 2016). A Poisson distribution assumes the dispersion parameter to be 1, and under/over-dispersion exists when the parameter is smaller/larger than 1 (Hardin & Hilbe, 2018). For the fixation count data, I would start with building a Poisson mixed effects model and then check the dispersion parameter with the performance package in R (version 0.4.8) (Lüdecke, Makowski, Waggoner, & Patil, 2020). If overdispersion was found, I would follow the eye-tracking literature in adopting the negative binomial distribution. As mentioned previously, only 0.36% (44) of the data were zero values, and since Poisson distribution and negative binomial distribution can handle non-excessive zero values (Zuur et al., 2009; Zuur & Ieno, 2016), I entered all data of fixation counts into building mixed effects models.

In summary, a generalized gamma distribution would be used for modeling fixation durations. For fixation counts, I would first try a Poisson distribution for building the mixed effects model and checking the dispersion parameter. If overdispersion was found, I would adopt a negative binomial distribution for building a new mixed effects model.

Fixed Effects and Random Effects. For the current RQ, presentation format was the

predictor of major interest. Additional predictors were also shared between fixation durations and fixation counts: element of a Chinese word (i.e., characters, pinyin, or English meaning) and time order of learning (i.e., first or second time). Random effects were assumed to come from both participant and word levels, and the maximal model (Barr et al., 2013) was used to specify the random effects and fixed effects structures in all models as I did in RQ 1.

Data Preparation. As mentioned previously, zero values in the data of fixation durations were converted to missing values from model building. Similar to RQ 1, categorical predictors (i.e., presentation format, time order and element) were recorded with the deviation coding scheme to assess main effects in the mixed effects models (Barr et al., 2013).

Generalized Mixed Effects Model for Fixation Durations. I used the glmer function from the lme4 package (version 1.1-23) (Bates et al., 2015) to build the generalized mixed effects model for fixation durations, by specifying the gamma family (a log link to ensure positive values), using the "bobyqa" optimizer (see Link & Cunnings, 2015), and setting the maximum iterations as 200,000 (see Miller, 2018). Similar to RQ 1, I started with the maximal model for both fixed effects and random effects, and the most parsimonious model was chosen based on the results of the anova function for model comparison based on AIC (Cunnings & Finlayson, 2015).

The formula for the final generalized mixed effects model for fixation durations was:

Dur ~ PF + Elem + Time + PF:Elem + Elem:Time + (1 + PF | ID) + (1 | Word)

In this formula, the outcome variable was non-zero fixation durations (Dur) at element level (i.e., characters, pinyin, or English meaning). The fixed effects were presentation format (PF; 3 levels), element (Elem; 3 levels), time order (Time; 2 levels), as well as two interactions between presentation format and element (PF:Elem) and between element and time order (Elem:

Time). The random effects structure consisted of a random intercept and a random slope of presentation format varied by participant (ID) and a random intercept varied by word (Word).

Table 41 shows the results of the final model. When considering presentation format alone, the mean fixation duration (on the gamma distribution scale) of the adjacent format was significantly shorter than the grant mean of fixation durations for all three presentation formats (Estimate = -0.029, $p = .005$), indicating that the adjacent format may lead to overall shorter fixation durations on all three elements, interestingly. In addition, a main effect was found for element: characters received significantly longer fixation durations (Estimate = 0.992, $p < .001$). A significant main effect was also found for time order: compared with the second time of learning, fixation durations were longer during the first time (Estimate = 0.014, $p = .005$). Interactions were also found between presentation format and element: in the adjacent format, characters received longer fixation durations (Estimate = 0.090, $p < .001$), whereas pinyin received shorter fixation durations (Estimate = -0.198, $p < .001$). Regarding time order, characters received significantly shorter fixation durations during the second time of learning (Estimate = -.020, $p = .007$).

To facilitate interpretation of the interactions, I drew two interaction plots based on the means of fixation durations (see Figures 30 & 31). As shown in Figure 30, the pattern of fixation durations of the adjacent format was clearly different from those of the other two formats, and the differences mainly lied in the data of characters and pinyin. The adjacent format promoted longer fixation durations on characters, but led to shorter fixation durations on pinyin. In Figure 31, the lines for pinyin and meaning were almost parallel in going downward from first to second time of learning, whereas the line for characters had the trend in moving upward, which showed an opposite pattern between characters and the other two elements.

Table 41. Results of Generalized Mixed Effects Model of Fixation Durations for RQ 2

Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		8.060*	0.023	347.458	< .001
PF-A		-0.029*	0.010	-2.816	.005
PF-H		0.040	0.010	3.935	< .001
Elem-C		0.992*	0.008	131.851	< .001
Elem-P		-0.161*	0.007	-21.783	< .001
Time-T1		0.014*	0.005	2.796	.005
PF-A:Elem-C		0.090*	0.011	8.507	< .001
PF-H:Elem-C		-0.083*	0.011	-7.852	< .001
PF-A:Elem-P		-0.198*	0.010	-19.043	< .001
PF-H:Elem-P		0.084*	0.010	8.049	< .001
Elem-C:Time-T1		-0.020*	0.007	-2.709	.007
Elem-P:Time-T1		0.013	0.008	1.805	.071
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.011	0.105		
	PF-A	0.002	0.042		
	PF-H	0.002	0.042		
Word	Intercept	0.001	0.028		

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Elem-C = Characters. Elem-P = Pinyin. Time-T1 = First Time of Learning.

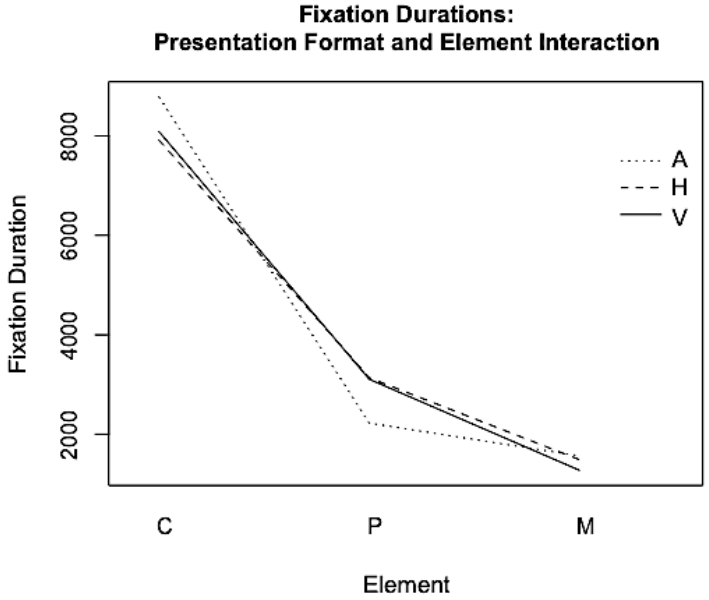

Figure 30. Interaction plot of presentation format and element interaction for fixation durations.

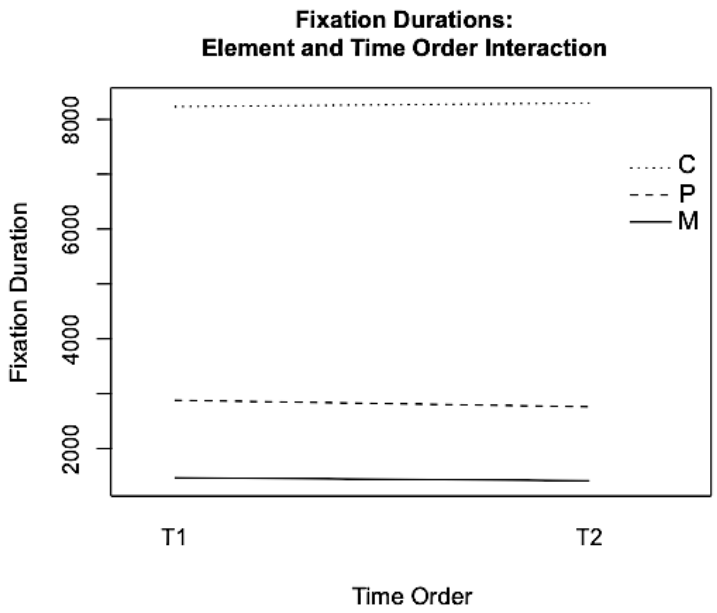

Figure 31. Interaction plot of presentation format and time order for fixation durations.

Combining the results from the generalized mixed effects model and the interaction plots, different presentation formats were found to result in different patterns of fixation durations on the three elements of Chinese words. Two different trends of fixation durations on characters against the other two elements were also found from the first to the second time of learning.

Generalized Mixed Effects Model for Fixation Counts. Given the needs to model count data with Poisson distribution and negative binomial distribution, I used glmmTMB (version 1.0.2.1) package (Brooks, Kristensen, van Benthem, Magnusson, Berg, Nielsen, Skaug, Mächler, & Bolker, 2017) to build the generalized mixed effects model for fixation counts. Same as lme4 (Bates et al., 2015), glmmTMB integrates random effects with maximum likelihood estimation and Laplace approximation, but offers more options of distribution families and an advantage of speed when estimating generalized linear mixed effects models (Brooke et al., 2017). It is also friendly to users of lme4 as glmmTMB shares the same syntax in writing the model formula with lme4 (Brooke et al., 2017). For model building, I started with finding the maximal model for random effects (Barr et al., 2013) while keeping all possible fixed effects

(i.e., main effects and interactions). After the maximal random effects structure was decided (i.e., the model converged), I selected the most parsimonious model in a similar way as I did for fixation durations. Specifically, I first checked for non-significant fixed effects in the model and performed model simplification by removing the non-significant fixed effects one by one. Then I used the anova function to compare the AICs between two models that differed only in the inclusion of fixed effects (Cunnings & Finlayson, 2015). If the AIC was significantly smaller in the simpler model, the simpler model was selected as the more parsimonious one.

As mentioned previously, I would first fit the data with a Poisson distribution (a log link) to calculate the dispersion parameter by using the performance package (version 0.4.8) (Lüdecke et al., 2020). Overdispersion was found (dispersion parameter = 2.127, $p < .001$), which is significantly larger than 1 (Hardin & Hilbe, 2018). Therefore, I proceeded with negative binomial distribution (nbinom1 with a log link; variance increases linearly, Brooke et al., 2017) to build the generalized mixed effects model.

The formula for the final generalized mixed effects model for fixation counts was:

Count ~ PF + Elem + Time + PF:Elem + Elem:Time + (1 | ID) + (1 | Word)

In this formula, the outcome variable was fixation counts (Count) at element level (i.e., characters, pinyin, or English meaning). The fixed effects were presentation format (PF; 3 levels), element (Elem; 3 levels), time order (Time; 2 levels), as well as two interactions between presentation format and element (PF:Elem) and between element and time order (Elem:Time). The random effects structure consisted of a random intercept varied by participant (ID) and a random intercept varied by word (Word).

Table 42 shows the results of the final model. Two main effects were found significant: element and time. Specifically, characters generally received more fixation counts (Estimate =

0.727, $p < .001$), whereas pinyin received fewer fixation counts (Estimate = -0.032, $p = .006$). Compared with the second time, there were more fixation counts during the first time of learning (Estimate = 0.051, $p < .001$). Two interactions were also found significant between presentation formant and element and between element and time. In the adjacent format, characters received more fixation counts (Estimate = 0.076, $p < .001$), while pinyin received fewer fixation counts (Estimate = -0.154, $p < .001$). Considering the time order, characters received more fixation counts during the second than the first time of learning (Estimate = -0.018, $p < .001$), whereas pinyin received fewer fixation counts during the second time (Estimate = -0.023, $p < .001$).

Table 42. Results of Generalized Mixed Effects Model of Fixation Counts for RQ 2

Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		2.303*	0.018	126.75	< .001
PF-A		-0.009	0.006	-1.410	.158
PF-H		0.005	0.006	0.81	.420
Elem-C		0.727*	0.005	135.89	< .001
Elem-P		-0.032*	0.006	-5.230	< .001
Time-T1		0.051*	0.004	11.640	< .001
PF-A:Elem-C		0.076*	0.007	10.140	< .001
PF-H:Elem-C		-0.036*	0.007	-4.760	< .001
PF-A:Elem-P		-0.154*	0.009	-17.540	< .001
PF-H:Elem-P		0.067*	0.009	7.780	< .001
Elem-C:Time-T1		-0.018*	0.005	-3.440	.001
Elem-P:Time-T1		0.023*	0.006	3.870	< .001
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.021	0.143		
Word	Intercept	< .001	0.012		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Elem-C = Characters. Elem-P = Pinyin. Time-T1 = First Time of Learning.

To facilitate interpretation of the interactions, I drew two interaction plots based on the means of fixation counts (see Figures 32 & 33). As shown in Figure 32, the adjacent format differed clearly from the other formats in the pattern of fixation counts on the three elements.

Specifically, more fixation counts occurred with characters and meaning in the adjacent format than the other two formats, whereas pinyin in the adjacent format received fewer fixation counts. In Figure 33, fixation counts were on a decreasing trend for all elements from the first to second time of learning, but the trend for meaning was slightly less intense than the other two elements.

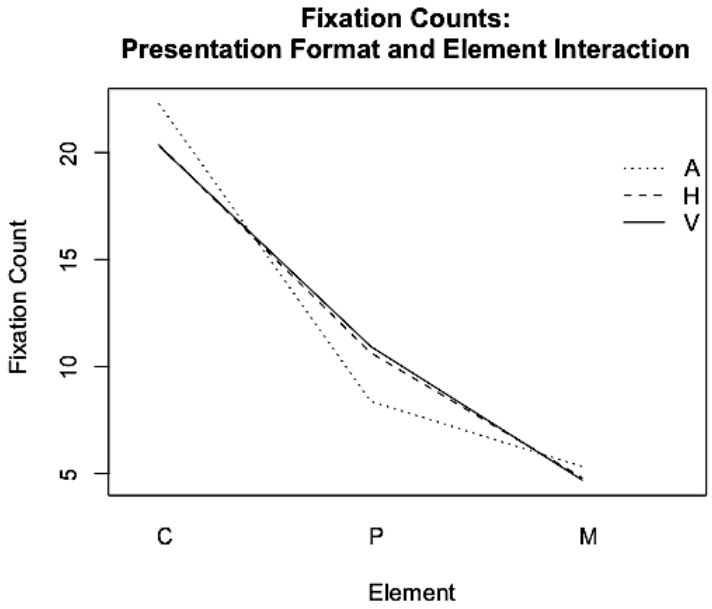

Figure 32. Interaction plot of presentation format and element for fixation counts.

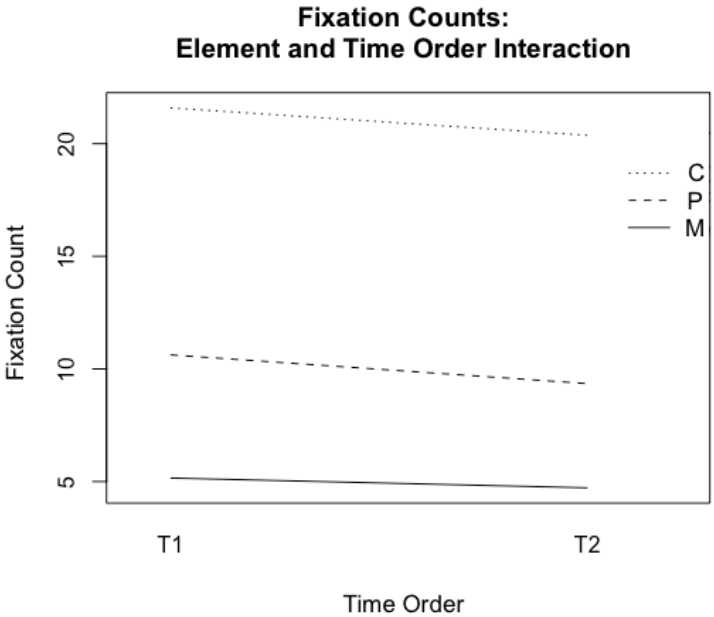

Figure 33. Interaction plot of element and time order for fixation counts.

Taking the results from the mixed effects model and the interaction plots together, different presentation formats were found to result in different patterns of fixation counts on the three elements of Chinese words, which was similar to the results of fixation durations. In addition, a common trend of fewer fixation counts from the first to the second time of learning was also found among the three elements of Chinese words, which was different from the results of fixation durations.

Table 43. Summary of Generalized Mixed Effects Models for RQ 2

Outcome Variable	Fixation Duration	Fixation Count
Fixed Effects	(–) PF^{*A}	(–) PF
	(+/–) Element*	(+/–) Element*
	(+) Time*	(+) Time*
	(+/–) PF:Element*	(+/–) PF:Element*
	(–) Element:Time^{*C}	(+/–) Element:Time*
Random Effects	(1 + PF | ID)	(1 | ID)
	(1 | Word)	(1 | Word)

Note. ^{*A} Statistically significant for a particular level of the categorical variable at $p < .05$: A = Adjacent format; C = Characters.
* Statistically significant (for more than one level of the categorical variable) at $p < .05$.
(+/–) indicates the sign of the estimate. PF = Presentation Format. Time = Time Order.

To compare the results between the two attention indices, I created a summary table (see Table 43) for the two generalized mixed effects models for fixation durations and fixation counts. As shown in Table 43, the fixed effects of the two models were very similar: main effects of presentation format, element, and time order, as well as an interaction between presentation format and element and between element and time order. These results confirmed that these three predictor variables played a role in directing learner attention during L2 Chinese vocabulary learning. Overall, the adjacent format led to clearly different patterns of the two attention indices for the three elements of Chinese words than the horizontal and vertical formats. Specifically, the adjacent format promoted more fixation durations and fixation counts on characters (and

meaning), while inevitably reduced those on the pinyin (and meaning). Regarding the effects of time order of learning, the overall trend was less fixation counts for the three elements. However, during the second time of learning, characters received longer fixation durations while the other two elements received shorter fixation durations.

4.5.3 RQ 3. What is the relationship between *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) and *learning outcomes* (as measured by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency and test formats into consideration?

Fixed Effects and Random Effects. This RQ can be regarded as an extension of RQ 1. The outcome variable was still vocabulary gain scores at the level of vocabulary test items. For the fixed effects, in addition to presentation formats (major interest), L2 Chinese proficiency, and vocabulary test format, the two attention indices (i.e., fixation durations and fixation counts) afforded new predictors of major interest. The random effects were still assumed to come from participants and words. As mentioned previously, the type of mixed effects model largely depends on the data distribution of the outcome variable (Cunnings & Finlayson, 2015; Gelman & Hill, 2007). Therefore, the two final models for RQ 1 (i.e., mixed logit model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl; two-part mixed effects model for M2C_rcl, C2P_rcl, and M2P_rcl) provided the baseline models for further model building in the current RQ.

Data Preparation. As mentioned previously, the learner attention data were collected from two times of learning for the three elements (i.e., characters, pinyin, and English meaning)

of Chinese words: fixation durations and fixation counts on each element. To investigate the overall impact of learner attention, I calculated the total values for each of the two attention indices by adding the data from the first and the second time of learning.

For each attention index, the data of the three elements of a Chinese word came from the same learning sessions that comprised the whole set of learning material for each word, so these data may inevitably correlate with each other. Results from bivariate correlations (see 4.2.1) confirmed the correlations between the attention data of the three elements. For fixation durations, all correlations between the three elements were statistically significant, and the size (i.e., regardless of the direction or the positive/negative sign of the coefficient) of these correlations ranged from 0.40 to 0.85. For fixation counts, all except two correlations between the three elements were statistically significant, with the size of the significant correlations ranging from 0.26 to 0.62. As correlations with a size larger than 0.70 signal potential issues of collinearity (Field, 2018), the attention data for the three elements of each word were very likely collinear, and when they were entered together as predictors into the mixed effects model, they may lead to confusing statistical analyses that fail to uncover statistically significant parameters (Zurr, Ieno, & Elphick, 2010).

As recommended by Zuur et al. (2010), principal component analysis (PCA) can identify and reduce collinearity by extracting fewer, distinctive components from a range of correlated variables, and by providing composite scores instead of original variable data in mixed effects modeling. However, when checking the directions/signs of the correlations between the elements for each attention index, some of them were at opposite directions. For fixation durations and fixation counts, the correlations between characters and the other two elements were negative, while the correlations between pinyin and English were positive (see 4.2.1). These correlations at

the opposite directions would increase the difficulty in interpreting the composite score calculated from PCA, because it is unclear whether one unit increased in the composite score was the remaining value after canceling out the opposite contributions from the three elements. Using composite scores of an omnibus nature may not help with detailing the relationships between attention to the *three elements* and vocabulary learning gains either. As a way to work around these issues, I decided to include the attention data of only one element each time for model building, instead of including all elements at the same time. Consequently, separate mixed effects models would be built for characters, pinyin, and English meaning of fixation durations and fixation counts, respectively.

As a general way to avoid collinearity among predictors and to increase interpretability of the results, the data of the two attention indices and L2 Chinese proficiency were standardized by following Gelman and Hill (2007) (see Figure 27 for the equation to calculate the *z* scores.). Same as in RQs 1 and 2, presentation formats and vocabulary test formats were recoded with the deviation coding scheme to assess main effects (Barr et al., 2013).

Model Building and Selection. I used the glmer function from the lme4 package (version 1.1-23) (Bates et al., 2015) for mixed effects modeling, by using the "bobyqa" optimizer (see Link & Cunnings, 2015), and setting the maximum iterations as 200,000 (see Miller, 2018). The two final models in RQ 1, namely, the mixed logit model (for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl) as well as the two-part mixed effects model (for M2C_rcl, C2P_rcl, and M2P_rcl) (see Table 44 for summary) provided the baseline models for model building in this RQ. For both fixed effects and random effects structures in the mixed effects models, I would start with the maximal model (Barr et al., 2013), for the current case, by adding the maximal structure involving the attention index to the baseline models. Specifically, the fixed

effects added to the baseline models would include the main effect of the attention index and its interactions with other main effects in the model (if applicable), and the random effects added would include a random slope of the attention index varied by participant and by word. After deciding the maximal random effects structure (i.e., the model converged), I would use the anova function in lme4 to compare the AICs of models that differed only in fixed effects (Cunnings & Finlayson, 2015), so as to select the most parsimonious model. Totally 12 mixed effects models were built for this RQ: 6 for fixation durations (2 for each element, #1-6) and 6 for fixation counts (2 for each element, #7-12) (see Appendix L for the formulae and results of each model).

Table 44. Summary of the Mixed Logit Model and the Two-Part Mixed Effects Model for RQ 1

Model	Mixed Logit	Two-Part Mixed Effects	
Outcome Variable	Gain (binary; 0 or 1)	Gain (fraction; between 0 and 1)	
Vocabulary Test Format	C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, C2M_rcl	M2C_rcl, C2P_rcl, M2P_rcl	
Fixed Effects	PF L2 Proficiency Test Format	Continuous Part	Binary Part
			L2 Proficiency Test Format
Random Effects	(1 + PF | ID) (1 + PF | Word)	(1 | ID) (1 | Word)	(1 | ID) (1 | Word)

Table 45 provides a summary of the fixed effects structures of the 12 mixed effects models for this RQ. For fixation durations (#1-3), the fixed effects structures for the five vocabulary test formats (i.e., C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl; mixed logit models) were very similar for characters, pinyin, and meaning: fixation durations, L2 Chinese proficiency, and vocabulary test formats as main effects, as well as an interaction between fixation durations and vocabulary test formats. Regarding the direction of effects (i.e., the sign of the estimate), fixation durations of characters offered a positive impact (i.e., longer fixation

durations led to higher learning gains), whereas those of pinyin and meaning had a negative impact (i.e., longer fixation durations led to lower learning gains).

Table 45. Summary of the Mixed Logit Models and the Two-Part Mixed Effects Models for RQ 3

Model	Mixed Logit	Two-Part	
Test Format	C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, C2M_rcl	M2C_rcl, C2P_rcl, M2P_rcl	
	#1. Fixation Durations: Characters		
Model	Mixed Logit	Two-Part: Continuous	Two-Part: Binary
Fixed Effects	(+) Prof*		(+) Prof*
	(+/–) TF*		(–) TF*
	(+) Dur.C*		(+) Dur.C
	(+/–) Dur.C:TF*		(–) Dur.C:TF*H
	#2. Fixation Durations: Pinyin		
Model	Mixed Logit	Two-Part: Continuous	Two-Part: Binary
Fixed Effects	(+) Prof*		(+) Prof*
	(+/–) TF*		(–) TF*
	(–) Dur.P*		(–) Dur.P*
	(+/–) Dur.P:TF*		(+) Dur.P:TF*H
	#3. Fixation Durations: Meaning		
Model	Mixed Logit	Two-Part: Continuous	Two-Part: Binary
Fixed Effects	(+) PF*A		
	(+) Prof*		(+) Prof*
	(+/–) TF*		(–) TF*
	(–) Dur.M*		(–) Dur.M*
	(+/–) Dur.M:TF*		
	#4. Fixation Counts: Characters		
Model	Mixed Logit	Two-Part: Continuous	Two-Part: Binary
Fixed Effects	(+) Prof*		(+) Prof*
	(+/–) TF*		(–) TF*
	(+) Count.C*		(+) Count.C*
	(+/–) Count.C:TF*		(–) Count.C:TF*H
	#5. Fixation Counts: Pinyin		
Model	Mixed Logit	Two-Part: Continuous	Two-Part: Binary
Fixed Effects	(+/–) PF		
	(+) Prof*		(+) Prof*
	(+/–) TF*		(–) TF*
	(–) Count.P*		
	(–) Count.P:PF*H		
	(+/–) Count.P:TF*		
	#6. Fixation Counts: Meaning		
Model	Mixed Logit	Two-Part: Continuous	Two-Part: Binary

Table 45 (cont'd)

Fixed Effects	(+) PF^{*A}	
	(+) Prof*	(+) Prof*
	(+/–) TF*	(–) TF*
	(–) Count.M*	(–) Count.M*
	(+/–) Count.M:TF*	

Note. ^{*A} Statistically significant for one particular level of the categorical variable at $p < .05$: A = Adjacent Format; H = Horizontal Format.
* Statistically significant (for more than one level of the categorical variable) at $p < .05$.
(+/–) indicates the sign of the estimate. Prof = Proficiency. TF = Vocabulary Test Format. Dur.C = Fixation Durations of Characters. Dur.P = Fixation Durations of Pinyin. Dur.M = Fixation Durations of Meaning. Count.C = Fixation Counts of Characters. Count.P = Fixation Counts of Pinyin. Count.M = Fixation Counts of Meaning.

For the other three vocabulary test formats (M2C_rcl, C2P_rcl, and M2P_rcl; two-part mixed effects models), the fixed effects of fixation durations for both the continuous part and the binary part were also similar for characters, pinyin, and meaning. Specifically, for the binary part, fixation durations, L2 Chinese vocabulary, and vocabulary test formats comprised main effects. As for the direction of effects in the binary part, fixation durations of characters offered a positive effect (i.e., the probability of obtaining learning gains increased as fixation durations became longer), whereas those of pinyin and meaning posted a negative effect (i.e., the probability of obtaining learning gains decreased for longer fixation durations). However, in the situation where learning gains were obtained, the amount of learning gains was not predicted by the predictors entered for mixed effects modeling (i.e., fixation durations, presentation formats, L2 Chinese proficiency, and vocabulary test formats), which may be caused by the low number (190) of non-zero item-level vocabulary gain scores available for model building (Zuur & Ieno, 2016). That is, such small amount of non-zero data (i.e., low variances in data) may not support complex modeling building and statistical analysis.

To summarize the results of fixation durations, the mixed logit models and the two-part

mixed effects models were very similar in the fixed effects structures: both types of models shared fixation durations, L2 Chinese proficiency, vocabulary test formats, and the interaction between fixation durations and vocabulary test formats (except for meaning) as main effects. The directions of effects (i.e., sign of the estimate) were also mostly shared between the two types of models, especially those of fixation durations. These results supported that learner attention as measured by fixation durations affected vocabulary learning gains that were assessed by the different eight test formats.

The results of fixation counts (#4-6) were very similar in pattern to those discussed for fixation durations. The fixed effects structures were mostly shared within the mixed logit models for the five test formats, within the binary parts of the two-part mixed effects models for the three test formats, and between these two types of models: fixation counts (except for pinyin), L2 Chinese proficiency, and vocabulary test formats. Additionally, the directions of the main effects were also mostly shared within the same type of models and between different types of models. Particularly, fixation counts of characters offered positive effects (i.e., more fixation counts resulted in higher probability of learning gains), whereas those of pinyin and meaning led to negative effects (i.e., more fixation durations led to lower probability of learning gains). For the continuous parts of the two-part mixed effects models, the lack of significant predictors may again be attributed to the small sample of non-zero vocabulary gain scores.

Considering the results of fixation durations and fixation counts together, both suggested learner attention indexed by fixation durations and fixation counts affected learning gains, and the direction of effects depended on what element (i.e., characters, pinyin, and meaning) learner paid attention to: more attention to characters led to better learning outcomes, whereas more attention to pinyin or meaning may not facilitate learning.

4.5.4 RQ 4. What is ***learners' preference*** (as measured by preference ratings) among the ***presentation formats*** (i.e., horizontal, vertical, and adjacent) in L2 Chinese vocabulary learning, taking their verbal reports into consideration?

For this RQ, the outcome variable was learners' preference ratings (based on a 7-point Likert scale), and the predictor was presentation format. Since three ratings came from the same participant, I conducted repeated-measures ANOVA for the data, using IBM SPSS Statistics 25. Bootstrapped descriptive statistics (see Table 30 in 4.1.4 Post-learning Survey) showed the preference ratings for each presentation format were generally normally distributed, so no further data transformation was needed.

Results showed that the assumption of sphericity was not violated (Mauchly's test of sphericity, W = 0.928, $p = .082$), but it is recommended to report the corrected results regardless of Mauchly's test results (Field, 2018). Therefore, below I reported the results based on the Greenhouse–Geisser correction, which is a stricter form of correction (Field, 2018). Significant difference was found among the preference ratings for the three presentation formats: Mean-Adjacent = 3.22, Mean-Horizontal = 5.23, Mean-Vertical = 4.48, $F = 14.796$, $p < .001$. Partial eta square was 0.179, which can be regarded as a large effect size according to Cohen's (1988) benchmark of 0.01 for small, 0.06 for medium, and 0.14 for large effect sizes. Pairwise comparisons showed that the adjacent format received significantly lower ratings than the horizontal or vertical formats did, whereas the ratings of the latter two did not differ significantly between themselves (see Table 46). These results indicated that the adjacent format was the least preferred by the participants. Verbal reports also echoed the current results, as although a few participants preferred the adjacent format over the other two formats, other participants were more likely to prefer the more commonly encountered horizontal and vertical formats.

Table 46. Results of Pairwise Comparisons Between Three Presentation Formats for RQ 4

				95% CI of Mean Difference	
		Mean Difference	*p*	*Lower*	*Upper*
Adjacent	Horizonal	2.01*	< .001	1.01	3.02
	Vertical	1.26*	.005	0.32	2.20
Horizontal	Vertical	0.75	.068	0.40	1.55

Note. *$p < .05$. Bonferroni correction has already been applied to the current results.

4.5.5 RQ 5. What is the relationship between *learners' preference* (as measured by preference ratings) and *learning outcomes* (as measured by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency, test formats, and their verbal reports into consideration?

Fixed Effects and Random Effects. This RQ can be regarded as an extension of RQ 1. The outcome variable was still vocabulary gain scores at the level of vocabulary test items. For the fixed effects, in addition to presentation formats (major interest), L2 Chinese proficiency, and vocabulary test format, the new predictor of major interest was preference ratings. The random effects were still assumed to come from participants and words. As mentioned previously, the type of mixed effects model largely depends on the data distribution of the outcome variable (Cunnings & Finlayson, 2015; Gelman & Hill, 2007). Therefore, the two final models for RQ 1 (i.e., mixed logit model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl; two-part mixed effects model for M2C_rcl, C2P_rcl, and M2P_rcl) provided the baseline models for further model building in the current RQ (see Table 44 in 4.5.3 RQ 3 for a summary of the two baseline models).

Model Building and Selection. As a general way to avoid collinearity among predictors and to increase interpretability of the results, preference ratings as well as L2 Chinese

proficiency scores were standardized by following Gelman and Hill (2007) (see Figure 27 for the equation to calculate the *z* scores.). Same as in previous mixed effects models in this study, presentation format and vocabulary test format were recoded with the deviation coding scheme to assess main effects (Barr et al., 2013).

I used the glmer function from the lme4 package (version 1.1-23) (Bates et al., 2015) for mixed effects modeling, by using the "bobyqa" optimizer (see Link & Cunnings, 2015), and setting the maximum iterations as 200,000 (see Miller, 2018). Same as previous model building in this study, for both fixed effects and random effects structures in the mixed effects models, I would start with the maximal model (Barr et al., 2013), for the current case, by adding the maximal structure involving preference ratings to the baseline models. Specifically, the fixed effects added to the baseline models would include the main effect of preference ratings and its interactions with other main effects in the model (if applicable), and the random effects added would include a random slope of preference ratings varied by participant. After deciding the maximal random effects structure (i.e., the model converged), I would use the anova function to compare the AICs of models that differed only in fixed effects (Cunnings & Finlayson, 2015), so as to select the most parsimonious model.

Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl. When the preference ratings were entered as a random slope varied by participant, the models failed to converge, so I kept the original random effects structure (i.e., a random intercept and a random slope of presentation format varied by participant and by word) for subsequent model building. Results showed the baseline model, whose fixed effects were presentation formats, L2 Chinese proficiency, and vocabulary test formats, did not improve significantly with the addition of preference ratings as a main effect or its interactions. No new mixed logit model was built for

Table 47. Results of Two-Part Mixed Effects Model for RQ 5

		Continuous				Binary			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.581*	0.086	-6.781	< .001	-5.446*	0.329	-16.542	< .001
Prof						0.811*	0.369	2.203	.028
M2C_rcl						-1.387*	0.243	-5.719	< .001
M2P_rcl						-0.544*	0.195	-2.793	.005
Pref						0.638*	0.299	2.131	.033
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.029	0.171			1.776	1.333		
	Pref					0.295	0.543		
Word	Intercept	0.028	0.142			0.926	0.962		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Pref = Preference Rating.

C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl.

Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl. The formulae for the final two-part mixed effects model were:

Continuous: Gain ~ 1 + (1 | ID) + (1 | Word)

Binary: Gain ~ Prof + TF + Pref + (1 + Pref | ID) + (1 | Word)

The continuous part stayed the same as in the baseline model: an intercept and a random effects structure consisting of a random intercept varied by participant and by word. For the binary part, the fixed effects were L2 Chinese proficiency (Prof; standardized), vocabulary test formats (TF; 3 levels), and preference ratings (Pref; standardized), and the random effects structure consisted of a random intercept and a random slope of preference ratings varied by participant (ID) and a random intercept varied by word (Word).

Table 47 shows the results of the two-part mixed effects model. Main effects were found significant for L2 Chinese proficiency, vocabulary test formats, and preference ratings. It was suggested that when the test formats were recall tasks of M2C_rcl, C2P_rcl, and M2P_rcl, preference ratings affected whether learning gains would be obtained or not: generally, the more a presentation format was preferred (i.e., higher preference ratings), the *more likely* learning gains would be obtained. When learning gains were obtained, its amount was not predicted by preference ratings, presentation formats, L2 Chinese proficiency, or vocabulary test formats.

Considering the results of all eight test formats, preference ratings did not predict whether learning gains were obtained or not for the five vocabulary test formats (i.e., C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl), but for the other three (i.e., M2C_rcl, C2P_rcl, and M2P_rcl), higher preference ratings led to more probability in obtaining learning gains.

4.5.6. RQ 6. What is the relationship between *learners' preference* (as measured by preference ratings) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking their verbal reports into consideration?

Fixed Effects and Random Effects. This RQ can be regarded as an extension of RQ 2. The outcome variable was still two attention indices: fixation durations and fixation counts. Specifically, the data of fixation durations and fixation counts were at element level (i.e., one data point for each element during each time of learning). For this RQ, preference rating, in addition to presentation format, was the predictor of major interest. Same as in RQ 2, fixation durations and fixation counts shared the same set of additional predictors: element of a Chinese word (i.e., characters, pinyin, or English meaning) and time order of learning (i.e., first or second time). Random effects were assumed to come from both participant and word levels, and the maximal model (Barr et al., 2013) was used to specify the random effects structure in all models.

Model Building and Selection. Since the type of mixed effects model largely depends on the data distribution of the outcome variable (Cunnings & Finlayson, 2015; Gelman & Hill, 2007), the two final generalized mixed effects models of RQ 2 (see Table 43 in 4.5.2 RQ 2 for a summary) provided the baseline models for further model building in the current RQ. Same as previous model building in this study, for both fixed effects and random effects structures in the mixed effects models, I would start with the maximal model (Barr et al., 2013), for the current case, by adding the maximal structure involving preference ratings to the baseline models. Specifically, the fixed effects added to the baseline models would include the main effect of preference ratings and its interactions with other main effects in the model (if applicable), and

the random effects added would include a random slope of preference ratings varied by participant. After deciding the maximal random effects (i.e., the model converged), I would use the anova function to compare the AICs of models that differed only in fixed effects (Cunnings & Finlayson, 2015), so as to select the most parsimonious model.

Data Preparation. Same as in RQ 2, zero values in the data of fixation durations were converted to missing values for model building. To avoid collinearity among predictors and to increase interpretability of the results, preference ratings were standardized by following Gelman and Hill (2007) (see Figure 27 for the equation to calculate the *z* scores.). Categorical predictors (i.e., presentation format, time order and element) were recorded with the deviation coding scheme to assess main effects in the mixed effects models (Barr et al., 2013).

Generalized Mixed Effects Model for Fixation Durations. I used the glmer function from the lme4 package (version 1.1-23) (Bates et al., 2015) to build the generalized mixed effects model for fixation durations, by specifying the gamma family (a log link to ensure positive values), using the "bobyqa" optimizer (see Link & Cunnings, 2015), and setting the maximum iterations as 200,000 (see Miller, 2018). The formula for the final model for fixation durations was:

Dur ~ PF + Elem + Time + Pref + PF:Elem + Elem:Time + Pref:Elem + (1 + Pref | ID) + (1 | Word)

In this formula, the outcome variable was non-zero fixation durations (Dur) at element level (i.e., characters, pinyin, or English meaning). The fixed effects were presentation format (PF; 3 levels), element (Elem; 3 levels), time order (Time; 2 levels), and preference ratings (Pref; standardized), as well as three interactions between presentation format and element (PF:Elem), between element and time order (Elem:Time), and between preference ratings and element

(Pref:Elem). The random effects structure consisted of a random intercept and a random slope of preference ratings varied by participant (ID) and a random intercept varied by word (Word).

Table 48 shows the results of the final model. Main effects of presentation format, element, and time order were found significant. Significant interactions were also found between presentation format and element, between characters and the first time of learning, and between pinyin and preference ratings. Specifically, for the interaction between pinyin and preference ratings, learners who gave a higher rating for a certain format would have shorter fixation durations on the pinyin (Estimate = -0.049, p = .001).

Table 48. Results of Generalized Mixed Effects Model of Fixation Durations for RQ 6

Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		8.060*	0.023	347.714	< .001
PF-A		-0.029*	0.008	-3.534	< .001
PF-H		0.040*	0.008	4.846	< .001
Elem-C		0.989*	0.008	131.583	< .001
Elem-P		-0.160*	0.007	-21.608	< .001
Time-T1		0.014*	0.005	2.731	.006
Pref		< .001	0.016	0.031	.975
PF-A:Elem-C		0.088*	0.010	8.491	< .001
PF-H:Elem-C		-0.082*	0.010	-7.895	< .001
PF-A:Elem-P		-0.197*	0.010	-18.999	< .001
PF-H:Elem-P		0.083*	0.010	8.030	< .001
Elem-C:Time-T1		-0.020*	0.007	-2.626	.009
Elem-P:Time-T1		0.013	0.007	1.813	.070
Elem-C:Pref		0.023	0.015	1.513	.130
Elem-P:Pref		-0.049*	0.015	-3.326	.001
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.011	0.104		
	Pref	0.003	0.059		
Word	Intercept	0.001	0.029		

Note. *p < .05.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Elem-C = Characters. Elem-P = Pinyin. Time-T1 = First Time of Learning. Prof = Proficiency. Pref = Preference Rating.

Generalized Mixed Effects Model for Fixation Counts. I used glmmTMB (version 1.0.2.1) (Brooks et al., 2017) to build the generalized mixed effects model for fixation counts. The formula for the final model for fixation counts was:

Count ~ PF + Elem + Time + Pref + PF:Elem + Elem:Time + Pref:PF + Pref:Elem + (1 | ID) + (1 | Word)

In this formula, the outcome variable was fixation counts (Count) at element level (i.e., characters, pinyin, or English meaning). The fixed effects were presentation format (PF; 3 levels), element (Elem; 3 levels), time order (Time; 2 levels), and preference ratings (Pref; standardized), as well as four interactions between presentation format and element (PF:Elem), between element and time order (Elem:Time), between preference rating and presentation format (Pref:PF), and between preference rating and element (Pref:Elem). The random effects structure consisted of a random intercept varied by participant (ID) and by word (Word).

Table 49. Results of Generalized Mixed Effects Model of Fixation Counts for RQ 6

Fixed Effects				
	Estimate	*Std.Error*	*z*	*p*
Intercept	2.303*	0.018	126.94	< .001
PF-A	-0.009	0.006	-1.38	.168
PF-H	0.005	0.006	0.75	.452
Elem-C	0.727*	0.005	135.910	< .001
Elem-P	-0.032*	0.006	-5.210	< .001
Time-T1	0.051*	0.004	11.630	< .001
Pref	0.014	0.009	1.540	.123
PF-A:Elem-C	0.076*	0.007	10.160	< .001
PF-H:Elem-C	-0.036*	0.007	-4.790	< .001
PF-A:Elem-P	-0.154*	0.009	-17.510	< .001
PF-H:Elem-P	0.067*	0.009	7.800	< .001
Elem-C:Time-T1	-0.018*	0.005	-3.430	.001
Elem-P:Time-T1	0.024*	0.006	3.86	< .001
PF-A:Pref	-0.020	0.015	-1.240	.214
PF-H:Pref	0.032*	0.015	2.200	.028
Elem-C:Pref	< -.001	0.011	-0.020	.983
Elem-P:Pref	-0.024*	0.012	-1.960	.050

Table 49 (cont'd)

Random Effects			
		Variance	*Std.Dev.*
ID	Intercept	0.021	0.143
Word	Intercept	< 0.001	0.012

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Elem-C = Characters. Elem-P = Pinyin. Time-T1 = First Time of Learning. Prof = Proficiency. Pref = Preference Rating.

Table 50. Summary of Generalized Mixed Effects Models of Fixation Durations and Fixation Counts for RQ 6

Outcome Variable	Fixation Duration	Fixation Count
Fixed Effects	(–) PF^{*A}	(–) PF
	(+/–) Element*	(+/–) Element*
	(+) Time*	(+) Time*
	(+) Pref	(+) Pref
	(+/–) PF:Element*	(+/–) PF:Element*
	(–) Element:Time^{*C}	(+/–) Element:Time*
		(–) PF:Pref^{*H}
	(–) Element:Pref^{*P}	(–) Element:Pref^{*P}
Random Effects	(1 + Pref | ID)	(1 | ID)
	(1 | Word)	(1 | Word)

Note. ^{*A} Statistically significant for a particular level of the categorical variable at $p < .05$: A = Adjacent; H = Horizontal; C = Characters; P = Pinyin.
* Statistically significant (for more than one level of the categorical variable) at $p < .05$.
(+/–) indicates the sign of the estimate. PF = Presentation Format. Time = Time Order. Pref = Preference Rating.

Table 49 shows the results of the final model. Main effects were found significant for element and time order, and interactions were also found significant between presentation format and element, between element and time order, between the horizontal format and preference ratings, and between pinyin and preference ratings. For the interaction between the horizontal format and preference ratings, those who gave the horizontal format higher ratings would have more fixation counts (Estimate = 0.032, $p = .028$). Regarding the interaction between pinyin and

preference ratings, those who gave higher ratings to a particular format would have fewer fixation counts on the pinyin.

To compare the results between the two attention indices, I created a summary table (see Table 50) for the two generalized mixed effects models for fixation durations and fixation counts. As shown in Table 50, the main effects were the same between two models: presentation format, element, time order, and preference rating. Regarding the interactions, the ones between presentation format and element, between element and time order, and between pinyin and preference ratings were shared between the two models. The model for fixation counts also had an interaction between the horizontal format and preference ratings. These results indicated that depending on the attention index, preference ratings interacted with a particular presentation or a specific element in affecting learner attention.

4.6 RQ Set B Focusing on Working Memory Capacities

4.6.1 RQ 7. What is the relationship between *working memory capacities* (as measured by a storage, a shifting, an updating, and an inhibition tasks) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning?

Fixed Effects and Random Effects. This RQ can be regarded as an extension of RQ 2. The outcome variable was still two attention indices: fixation durations and fixation counts. Specifically, the data of fixation durations and fixation counts were at element level (i.e., one data point for each element during each time of learning). For this RQ, working memory capacities, in addition to presentation format, was the predictor of major interest. Same as in RQ

2, the same set of additional predictors were also included for fixation durations and fixation counts: element of a Chinese word (i.e., characters, pinyin, or English meaning) and time order of learning (i.e., first or second time). Random effects were assumed to come from both participant and word levels.

Data Preparation. Same as in RQ 2, zero values in the data of fixation durations were converted to missing values for model building. The composite scores of four working memory tasks (see 4.1.3 Working Memory Tasks) were used for model building. Categorical predictors (i.e., presentation format, time order, and element) were recorded with the deviation coding scheme to assess main effects in the mixed effects models (Barr et al., 2013).

Model Building and Selection. Since the type of mixed effects model largely depends on the data distribution of the outcome variable (Cunnings & Finlayson, 2015; Gelman & Hill, 2007), the two final generalized mixed effects models of RQ 2 (see Table 43 in 4.5.2 RQ 2 for a summary) provided the baseline models for further model building in the current RQ. Same as previous model building in this study, for both fixed effects and random effects structures in the mixed effects models, I would start with the maximal model (Barr et al., 2013), for the current case, by adding the maximal structure involving working memory composite scores to the baseline models. Specifically, the fixed effects added to the baseline models would include a main effect of working memory composite scores and its interactions with other main effects in the model (if applicable). Since working memory capacities is a between-subject variable (i.e., each participant had one composite score), no additional component would be added to the random effects structure. After deciding the maximal random effects (i.e., the model converged), I would use the anova function to compare the AICs of models that differed only in fixed effects (Cunnings & Finlayson, 2015), so as to select the most parsimonious model.

Generalized Mixed Effects Model for Fixation Durations. I used the glmer function from the lme4 package (version 1.1-23) (Bates et al., 2015) to build the generalized mixed effects model for fixation durations, by specifying the gamma family (a log link to ensure positive values), using the "bobyqa" optimizer (see Link & Cunnings, 2015), and setting the maximum iterations as 200,000 (see Miller, 2018). The formula for the final generalized mixed effects model for fixation durations was:

Dur ~ PF + Elem + Time + WM + PF:Elem + Elem:Time + WM:Elem + (1 + PF | ID) + (1 | Word)

In this formula, the outcome variable was non-zero fixation durations (Dur) at element level (i.e., characters, pinyin, or English meaning). The fixed effects were presentation format (PF; 3 levels), element (Elem; 3 levels), time order (Time; 2 levels), and working memory capacities (WM; composite scores), as well as three interactions between presentation format and element (PF:Elem), between element and time order (Elem:Time), and between working memory capacities and element (WM:Elem). The random effects structure consisted of a random intercept and a random slope of presentation format varied by participant (ID) and a random intercept varied by word (Word).

Table 51 shows the results of the final model. Main effects of presentation format, element, and time order were found significant. Significant interactions were found between presentation format and element, between characters and the first time of learning, and between elements and working memory capacities. Specifically, for the interaction between elements and working memory capacities, learners with higher working memory capacities had longer fixation durations on characters, but shorter fixation durations on pinyin. These results indicated that working memory capacities affected learner attention (as indexed by fixation durations) to

different elements in different ways.

Table 51. Results of Generalized Mixed Effects Model of Fixation Durations for RQ 7

Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		8.060*	0.023	354.399	< .001
PF-A		-0.029*	0.010	-2.812	.005
PF-H		0.040*	0.011	3.925	< .001
Elem-C		0.989*	0.007	132.339	< .001
Elem-P		-0.160*	0.007	-21.748	< .001
Time-T1		0.014*	0.005	2.908	.004
WM		-0.039	0.021	-1.836	.067
PF-A:Elem-C		0.090*	0.010	8.578	< .001
PF-H:Elem-C		-0.083*	0.011	-7.700	< .001
PF-A:Elem-P		-0.199*	0.010	-19.137	< .001
PF-H:Elem-P		0.083*	0.010	7.977	< .001
Elem-C:Time-T1		-0.020*	0.007	-2.750	.006
Elem-P:Time-T1		0.014	0.007	1.921	.055
Elem-C:WM		0.088*	0.007	11.722	< .001
Elem-P:WM		-0.066*	0.008	-8.810	< .001
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.011	0.103		
	PF-A	0.002	0.042		
	PF-H	0.002	0.042		
Word	Intercept	< 0.001	0.029		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Elem-C = Characters. Elem-P = Pinyin. Time-T1 = First Time of Learning. Prof = Proficiency. WM = Working Memory Composite Score.

Generalized Mixed Effects Model for Fixation Counts. I used glmmTMB (version 1.0.2.1) (Brooks et al., 2017) to build the generalized mixed effects model for fixation counts. The formula for the final model for fixation counts was:

Count ~ PF + Elem + Time + WM + PF:Elem + Elem:Time + WM:Elem + (1 | ID) + (1 | Word)

In this formula, the outcome variable was fixation counts (Count) at element level (i.e., characters, pinyin, or English meaning). The fixed effects were presentation format (PF; 3

levels), element (Elem; 3 levels), time order (Time; 2 levels), and working memory capacities (WM; composite scores), as well as three interactions between presentation format and element (PF:Elem), between element and time order (Elem:Time), and between working memory capacities and element (WM:Elem). The random effects structure consisted of a random intercept varied by participant (ID) and a random intercept varied by word (Word).

Table 52. Results of Generalized Mixed Effects Model of Fixation Counts for RQ 7

Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		2.302*	0.018	129.380	< .001
PF-A		-0.009	0.006	-1.430	.153
PF-H		0.005	0.006	0.810	.416
Elem-C		0.725*	0.005	136.21	< .001
Elem-P		-0.030*	0.006	-4.870	< .001
Time-T1		0.052*	0.004	11.660	< .001
WM		0.007	0.018	0.410	.681
PF-A:Elem-C		0.075*	0.007	10.180	< .001
PF-H:Elem-C		-0.036*	0.007	-4.790	< .001
PF-A:Elem-P		-0.155*	0.009	-17.680	< .001
PF-H:Elem-P		0.067*	0.008	7.880	< .001
Elem-C:Time-T1		-0.018*	0.005	-3.460	.001
Elem-P:Time-T1		0.023*	0.006	3.830	< .001
Elem-C:WM		0.070*	0.005	12.970	< .001
Elem-P:WM		-0.053*	0.006	-8.530	< .001
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.020	0.141		
Word	Intercept	< 0.001	0.012		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Elem-C = Characters. Elem-P = Pinyin. Time-T1 = First Time of Learning. Prof = Proficiency. WM = Working Memory Composite Scores.

Table 52 shows the results of the final model. Main effects were found significant for element and time order, and interactions were also found significant between presentation format and element, between element and time order, and between elements and working memory capacities. For the interactions between elements and working memory capacities, those who had

higher working memory capacities would have more fixation counts on characters but fewer fixation counts on pinyin. These results indicated that working memory capacities affected learner attention (as indexed by fixation counts) to different elements in different ways.

To compare the results between the two attention indices, I created a summary table (see Table 53) for the two generalized mixed effects models for fixation durations and fixation counts. As shown in Table 53, the main effects were the same between two models: presentation format, element, time order, and working memory capacities. Three interactions were also shared between the two models: between presentation format and element, between element and time order, and between working memory capacities and element. These results indicated that working memory capacities affected learner attention to different elements in different ways.

Table 53. Summary of Generalized Mixed Effects Models of Fixation Durations and Fixation Counts for RQ 7

Outcome Variable	Fixation Duration	Fixation Count
Fixed Effects	(+/–) PF*	(+/–) PF
	(+/–) Element*	(+/–) Element*
	(+) Time*	(+) Time*
	(–) WM	(+) WM
	(+/–) PF:Element*	(+/–) PF:Element*
	(–) Element:Time*C	(+/–) Element:Time*
	(+/–) Element:WM*	(+/–) Element:WM*
Random Effects	(1 + PF | ID)	(1 | ID)
	(1 | Word)	(1 | Word)

Note. *C Statistically significant for a particular level of the categorical variable at $p < .05$: C = Characters.
* Statistically significant (for more than two levels of the categorical variable) at $p < .05$.
(+/–) indicates the sign of the estimate. PF = Presentation Format. Time = Time Order. WM = Working Memory Composite Score.

4.6.2 RQ 8. What is the relationship between *working memory capacities* (as measured by a storage, a shifting, an updating, and an inhibition tasks) and *learning outcomes* (as assessed by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency and test formats into consideration?

Fixed Effects and Random Effects. This RQ can be regarded as an extension of RQ 1. The outcome variable was still vocabulary gain scores at the level of vocabulary test items. For the fixed effects, in addition to presentation formats (major interest), L2 Chinese proficiency, and vocabulary test format, the new predictor of major interest was working memory capacities. The random effects were still assumed to come from participants and words. As mentioned previously, the type of mixed effects model largely depends on the data distribution of the outcome variable (Cunnings & Finlayson, 2015; Gelman & Hill, 2007). Therefore, the two final models for RQ 1 (i.e., mixed logit model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl; two-part mixed effects model for M2C_rcl, C2P_rcl, and M2P_rcl) provided the baseline models for further model building in the current RQ (see Table 44 in 4.5.3 RQ 3 for a summary of the two baseline models).

Data Preparation. To avoid collinearity among predictors and to increase interpretability of the results, L2 Chinese proficiency scores were standardized by following Gelman and Hill (2007) (see Figure 27 for the equation to calculate the *z* scores.). The composite scores of four working memory tasks (see 4.1.3 Working Memory Tasks) were used for model building. Same as in previous mixed effects modeling in this study, presentation format was recoded with the deviation coding scheme to assess main effects (Barr et al., 2013).

Model Building and Selection. I used the glmer function from the lme4 package

(version 1.1-23) (Bates et al., 2015) for mixed effects modeling, by using the "bobyqa" optimizer (see Link & Cunnings, 2015), and setting the maximum iterations as 200,000 (see Miller, 2018). Same as previous model building in this study, for both fixed effects and random effects structures in the mixed effects models, I would start with the maximal model (Barr et al., 2013), for the current case, by adding the maximal structure involving working memory composite scores to the baseline models. Specifically, the fixed effects added to the baseline models would include the main effect of working memory composite scores and its interactions with other main effects in the model (if applicable). Since working memory capacities is a between-subject variable (i.e., each participant had one composite score), no additional component would be added to the random effects structure. After deciding the maximal random effects (i.e., the model converged), I would use the anova function to compare the AICs of models that differed only in fixed effects (Cunnings & Finlayson, 2015), so as to select the most parsimonious model.

Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl. The formula for the final mixed logit model was:

Gain ~ Prof + TF + WM + WM:TF + (1 + PF | ID) + (1 + PF | Word)

In the formula, the outcome variable was item-level vocabulary gain scores (Gain; 0 or 1), and the fixed effects were L2 Chinese proficiency (Prof; standardized), vocabulary test formats (TF; 5 levels), and working memory capacities (WM; composite scores), as well as an interaction between working memory capacities and vocabulary test formats (WM:TF). The random effects structure consisted of a random intercept and a random slope of presentation format varied by participant (ID) and by word (Word) respectively.

Table 54 shows the results of the final model. All main effects were found significant: L2 Chinese proficiency, vocabulary test format, and working memory capacities. The interactions

between working memory capacities and vocabulary test formats were also significant. Specifically, for recognition test formats of M2C_rcg and C2P_rcg as well as recall test format of C2M_rcl, the higher working memory capacities, the more likely to obtain learning gains (i.e., the estimates were positive). Differently, for the recognition test format of M2P_rcg, higher working memory capacities were associated with less probability of obtaining learning gains (i.e., the estimate was negative).

Table 54. Results of Mixed Logit Model for RQ 8

Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.588*	0.169	-3.476	.001
Prof		0.667*	0.237	2.818	.005
M2C_rcg		0.842*	0.047	17.891	< .001
M2P_rcg		-0.227*	0.047	-4.831	< .001
C2M_rcl		-1.045*	0.054	-19.205	< .001
C2M_rcg		1.001*	0.048	20.952	< .001
WM		0.375*	0.115	3.255	.001
M2C_rcg:WM		0.112*	0.048	2.316	.021
M2P_rcg:WM		-0.355*	0.047	-7.532	< .001
C2M_rcl:WM		0.288*	0.056	5.132	< .001
C2M_rcg:WM		0.158*	0.049	3.234	.001
Random Effects					
		Variance	*Std.Dev.*		
ID	(Intercept)	0.835	0.914		
	PF-A	0.086	0.293		
	PF-H	0.075	0.275		
Word	(Intercept)	0.424	0.651		
	PF-A	0.031	0.176		
	PF-H	0.033	0.181		

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = L2 Chinese Proficiency.

Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl. The formulae for the final two-part mixed effects model were:

Continuous: Gain ~ 1 + (1 | ID) + (1 | Word)

Binary: Gain ~ TF + WM + WM:TF + (1 | ID) + (1 | Word)

For both parts, the outcome variable was item-level vocabulary gain scores (Gain; fraction score between 0 and 1). Regarding the continuous part, no fixed effects were found to be significant, and the random effects structure consisted of a random intercept varied by participant (ID) and by word (Word) respectively. In terms of the binary part, the fixed effects were vocabulary test formats (TF; 3 levels) and working memory capacities (WM; composite scores), as well as an interaction between working memory capacities and vocabulary test formats (WM:TF). The random effects structure consisted of a random intercept varied by participant (ID) and by word (Word) respectively.

Table 55 shows the results of the final model. For the continuous part, only the intercept was found as significant in the fixed effects component. This indicated that when there were vocabulary gains, the extent of gains (i.e., exact gain scores) were not associated with presentation formats, L2 Chinese proficiency, vocabulary test formats, or working memory capacities. As for the binary part, when considering whether vocabulary gains would be obtained or not, vocabulary test formats as well as the interaction between recall test format M2C_rcl and working memory capacities were found as significant predictors. Specifically, for recall test format M2P_rcl, higher working memory capacities resulted in lower probability of obtaining learning gains. These results indicated that working memory capacities affected the probability of obtaining learning gains assessed via the recall test format M2P_rcl.

Summary of Mixed Effects Models for RQ 8. Table 56 summarizes the two mixed effects models for all vocabulary test formats. Results showed that the fixed effects structure of the mixed logit model (for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl) was similar to that of the binary part of the two-part mixed effects model: vocabulary test formats, working

memory capacities, and their interaction. The directions of the effects (i.e., signs of the estimates) of these predictors were also similar. Generally, working memory capacities had a positive effect on the probability of obtaining learning gains: learners with higher working memory capacities were more likely to answer the vocabulary test item correctly. However, the advantage of higher working memory capacities could turn to disadvantage for some particular vocabulary test formats.

Table 55. Results of Two-Part Mixed Effects Model for RQ 8

		Continuous				Binary			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.581*	0.086	-6.781	< .001	-5.509*	0.336	-16.404	< .001
M2C_rcl						-1.480*	0.292	-5.062	< .001
M2P_rcl						-0.524*	0.223	-2.353	.019
WM						0.163	0.229	0.714	.476
M2C_rcl:WM						0.511	0.262	1.950	.051
M2P_rcl:WM						-0.766*	0.210	-3.646	< .001
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.029	0.171			1.767	1.329		
Word	Intercept	0.020	0.142			0.947	0.973		

Note. *$p < .05$. *Std.Error* = Standard Error. *Std.Dev.* = Standard Deviation. WM = Working Memory Composite Score.

Table 56. Summary of the Mixed Logit Models and the Two-Part Mixed Effects Models for RQ 8

Model	Mixed Logit	Two-Part: Continuous	Two-Part: Binary
Test Format	C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, C2M_rcl	M2C_rcl, C2P_rcl, M2P_rcl	
Fixed Effects	(+) Prof*		
	(+/–) TF*		(–) TF*
	(+) WM*		(+) WM
	(+/–) WM:TF*		(–) WM:TF*M2P
Random Effects	(1 + PF | ID)	(1 | ID)	(1 | ID)
	(1 + PF | Word)	(1 | Word)	(1 | Word)

Note. *M2P Statistically significant for recall test format M2P_rcl at $p < .05$:

* Statistically significant (for more than one level of the categorical variable) at $p < .05$.

(+/–) indicates the sign of the estimate. Prof = Proficiency. TF = Test Format. WM = Working Memory Composite Score.

CHAPTER 5 DISCUSSION AND CONCLUSION

In this chapter, I will first discuss the results for each RQ by referring to the current theories and research findings. Lastly in the Conclusion, I will discuss the results of all RQs more generally and offer pedagogical implications based on the current findings.

5.1 RQ Set A Focusing on Presentation Formats

5.1.1 RQ 1. What is the relationship between ***presentation formats*** (i.e., horizontal, vertical, and adjacent) and ***learning outcomes*** (as assessed by a bilingual vocabulary test) in L2 Chinese vocabulary learning, taking L2 Chinese proficiency and test formats into consideration?

HP 1: When L2 Chinese proficiency and vocabulary test formats are taken into consideration, the adjacent format will be associated with the highest gain scores, followed by the vertical and the horizontal format.

Results from the mixed effects model for the five vocabulary test formats (i.e., M2C_rcg, M2P_rcg, C2M_rcg, C2P_rcg, and C2M_rcl) showed that compared with the horizontal and vertical formats, the adjacent format were slightly more likely to result in learning gains. For the other three vocabulary test formats (i.e., M2C_rcl, M2P_rcl, and C2P_rcl), presentation format was not found as a significant contributor. Results from descriptive statistics (see 4.1.1 Vocabulary Pretest and Posttest) also supported the benefits of the adjacent format for better learning outcome.

The evidence that favored the adjacent format may seem not strong at a first glance. However, the participants in the current study had only 34 seconds to study a two-character Chinese word that was barely known, so the magnitude of effects could be expected to be

relatively small, especially that learning Chinese words has been well recognized as a major challenge for English native speakers. Another consideration came from the preference ratings. It was clear that most participants liked the adjacent format the least, and data from the verbal reports showed that they felt the adjacent format was "all over the place" and they did not know "where to look at." The novelty of the adjacent format and the uncomfortableness associated it may have prevented the participants from fully enjoying the benefits of the adjacent format. In real classrooms where students can receive more guidance about how to direct their attention, the advantages of the adjacent format may become more apparent with longer study time.

The current results were generally in accordance with the hypothesis that when L2 Chinese proficiency and vocabulary test formats were taken into consideration, the adjacent format would facilitate vocabulary learning. Notably, the five vocabulary test formats generally received higher scores (see 4.1.1 Vocabulary Pretest and Posttest) than the other three vocabulary test formats, which indicated that vocabulary knowledge assessed by these five vocabulary test formats was better developed at the initial stage of L2 Chinese vocabulary learning. The current results also supported Lee and Kalyuga's (2011) speculation that the adjacent format would lead to better learning outcome of L2 Chinese words. However, by including eight different vocabulary test formats, the current study went beyond their focus on recognition tasks and examined the development of vocabulary knowledge at more fine-grained stages.

Lastly, L2 Chinese proficiency was found as a significant predictor in the mixed effects models, which was also supported by the results of bivariate correlations (see 4.2.3 Vocabulary Gain Scores and L2 Chinese Proficiency). These results indicated that the effects of presentation formats should be considered along with L2 proficiency. Previous studies have also suggested

the role of L2 proficiency in mediating the effects of presentation formats. For example, in Yeung et al. (1997, 1999), students at higher proficiency levels tended to perform better in vocabulary tests than those at lower proficiency levels when working on an integrated format of a reading passage with vocabulary definitions. In the current study, higher L2 Chinese proficiency levels generally facilitated learning.

5.1.2 RQ 2. What is the relationship between *presentation formats* (i.e., horizontal, vertical, and adjacent) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in L2 Chinese vocabulary learning?

HP 2: Each presentation format will be associated with a different pattern of data for the attention indices of characters, pinyin, and meaning, i.e., a different combination of large and small numbers for the three attention indices of the three elements in a particular presentation format. Within each presentation format, characters will receive the largest numbers of fixation durations and fixation counts.

The current results showed that the adjacent format led to a clearly different pattern of fixation durations and fixation counts on the characters, pinyin, and English meaning, compared with the similar patterns between the horizontal and vertical formats. Interestingly, the adjacent format as a main effect led to overall lower total fixation durations than the other two formats. However, when fixation durations were broken down to the three elements, the adjacent format promoted longer fixation durations on characters and more fixation counts on characters and meaning than the other two formats. Meanwhile, the adjacent format also led to shorter fixation durations on pinyin and meaning, and fewer fixation counts on pinyin. These results were generally similar to those from the descriptive statistics (see 4.1.2 Eye-tracking). Therefore, the

hypothesis was generally supported: the patterns of attention were different across the three presentation formats, and the adjacent format promoted more attention to characters (and meaning) but meanwhile leaving less attention to pinyin (and meaning).

5.1.3 RQ 3. What is the relationship between ***learner attention*** (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) and ***learning outcomes*** (as measured by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency and test formats into consideration?

HP 3: When L2 Chinese proficiency and vocabulary test formats are taken into consideration, larger numbers of overall fixation durations and fixation counts will be associated with higher vocabulary scores in the three presentation formats.

The results from mixed effects modeling showed that generally across the three presentation formats, fixation durations and fixation counts could affect vocabulary learning gains in a positive or negative way. Specifically, longer fixation durations and more fixation counts on characters led to better learning outcome, whereas longer fixation durations and more fixation counts on pinyin and meaning had a negative impact on learning outcome. Notably, with different vocabulary test formats, the positive effects of more fixation durations/counts on learning outcome may reverse to negative effects.

These results suggested that when learning L2 Chinese words, despite pinyin and meaning as available information for learning, focusing on the characters would enhance overall mastery of the Chinese words. Following this reasoning, results from RQ 2 that the adjacent format promoted particularly longer fixation durations and more fixation counts on characters

but less so on pinyin or English meaning, would favor the adjacent format as superior to the other two formats. That is, the adjacent format increased fixation durations and fixation counts on characters, and because increased fixation durations and fixation counts on characters generally led to better learning outcome, the adjacent format led to better learning outcome as shown in RQ 1.

The overall positive effects of fixation durations and fixation counts on characters in this study generally echoed previous eye-tracking studies on L2 vocabulary (e.g., Godfroid et al., 2013) and L2 grammar learning (e.g., Indrarathne & Kormos, 2017) that more attention led to better learning outcome. However, the current findings showed that not only the amount of attention but also the focus of attention was important in learning: focusing too much on less substantial information may hinder learning. Results from bivariate correlations between vocabulary gain scores and fixation durations and fixation counts (see 4.2.2 Vocabulary Gain Scores and Fixation Durations/Counts) generally confirmed the results from the mixed effects modeling.

To summarize, the current results partially supported the hypothesis. Specifically, the overall facilitation of fixation durations and fixation counts to learning outcome mainly came from characters but not pinyin or meaning.

5.1.4 RQ 4. What is ***learners' preference*** (as measured by preference ratings) among the ***presentation formats*** (i.e., horizontal, vertical, and adjacent) in L2 Chinese vocabulary learning, taking their verbal reports into consideration?

HP 4: When learners' verbal reports are taken into consideration, the adjacent format will have the highest preference ratings, followed by the vertical and the adjacent format.

Contrary to the hypothesis, results from mixed effects modeling showed the adjacent format received significantly lower ratings than the horizontal and vertical formats. Descriptive statistics (see 4.1.4 Post-learning Survey) also offered similar results. Additionally, data from verbal reports confirmed the overall low preference on the adjacent format. One common reason from the participants who did not like the adjacent format was that they felt "it was all over the place" and they did not know "where to look at." However, a few participants preferred the adjacent format over the other two formats, as they felt everything was in their "peripheral vision" and they did not need to move their eyes around.

The horizontal format was the most commonly preferred one, and one major reason was its familiarity to the participants, as they explained that was how the textbooks and their notes were presented. They felt it more natural and easier to follow. The vertical format was kind of in the middle between the most liked horizontal format and the least preferred adjacent format. The reasons for liking included that it was similar to the horizonal format and that things were still in the same order. The reasons for disliking could be the difficulty in moving the eyes up and down.

The overall low preference of the adjacent format may explain its marginal significance as a contributor to better learning outcome as shown in RQ 1. The uncomfortableness most of the participants felt with the adjacent format may have prevented them from fully enjoying the benefits offered by it. Better learning outcome may be obtained as participants became more familiar with the adjacent format.

5.1.5 RQ 5. What is the relationship between ***learners' preference*** (as measured by preference ratings) and ***learning outcomes*** (as measured by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency, test formats, and their verbal reports into consideration?

HP 5: When L2 Chinese proficiency, vocabulary test formats, and learners' verbal reports are taken into consideration, higher preference ratings will be associated with higher vocabulary scores.

The current results from mixed effects modeling partially supported the hypothesis, as the positive relationship between preference ratings and learning outcome was only observed for the three vocabulary test formats of M2P_rcl, M2C_rcl, and C2P_rcl. Specifically, higher preference ratings were associated with more probability of answering correctly in the three vocabulary test formats. However, results from bivariate correlations did not share similar results (see 4.2.6 Vocabulary Gain Scores and Preference Ratings). This discrepancy may be because the bivariate correlations were conducted for the scores of all vocabulary test formats, and the aggregated scores may not reveal the relationships between learning outcome and preference ratings for three out of eight vocabulary test formats. These findings suggested that for difficult tasks such as recalling the pinyin or characters, choosing a preferred format would be more helpful, but for easier tasks, such as recognition tasks or recalling the meaning, choosing a preferred presentation format may not matter much. Overall, preference among the three presentation formats may not affect learning outcome in a significant way.

5.1.6. RQ 6. What is the relationship between *learners' preference* (as measured by preference ratings) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking their verbal reports into consideration?

HP 6: When learners' verbal reports are taken into consideration, higher preference ratings will be associated with larger numbers of overall fixation durations and fixation counts,.

Results from mixed effects modeling showed that the effects of preference ratings on attention depended on the elements and the presentation formats. Specifically, higher preference ratings led to shorter fixation durations and fewer fixation counts to pinyin. Overall, the effects of preference among the three presentation formats on attention can be regarded as not substantial, given the sparseness of the significant relationships found in the analysis. Therefore, the hypothesis was not fully supported by the current results.

5.2 RQ Set B Focusing on Working Memory Capacities

5.2.1 RQ 7. What is the relationship between *working memory capacities* (as measured by a storage, a shifting, an updating, and an inhibition tasks) and *learner attention* (to characters, pinyin, and meaning as indexed by fixation durations and fixation counts) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning?

HP 7: Higher working memory capacities will be associated with larger numbers of overall fixation durations and fixation counts.

Results from mixed effects modeling showed that depending on the element, working

memory capacities affected fixation durations and fixation counts in different ways. Specifically, for fixation durations, higher working memory capacities led to longer fixation durations on characters but shorter fixation durations on pinyin. Fixation counts also shared the same pattern of positive effects of working memory capacities on characters and negative effects on pinyin. Results from descriptive statistics partially confirmed the relationships between working memory capacities and attention indices: negative correlations with pinyin for fixation durations and positive correlations with characters for fixation counts. The discrepancy between the results of correlations and mixed effects modeling may lie in the difference of how they handle other related factors: whereas correlations may not account for the effects from other factors other than the two in the correlations, mixed effects modeling considered multiple factors together. The current results partially supported the hypothesis.

The effects of working memory capacities on attention was consistent with previous findings on L2 grammar learning (Indrarathne & Kormos, 2017). However, the relationship between working memory capacities and attention was more complex in the current study: the direction of effects of working memory capacities depended on the element to which learners paid attention. The finding that learners with higher working memory capacities paid more attention to characters but less to pinyin indicated a strategy of attention allocation supported by higher working memory capacities. That is, because of higher working memory capacities, learners would not need to spend much time on the pinyin and would be able to focus more on the characters. As results in RQ 3 showed that more attention to characters generally led to better overall learning outcome, such a strategy of focusing on characters also gave learners with higher working memory capacities an advantage in obtaining overall learning gains, which will be discussed in 5.2.2 RQ 8.

5.2.2 RQ 8. What is the relationship between *working memory capacities* (as measured by a storage, a shifting, an updating, and an inhibition tasks) and *learning outcomes* (as assessed by a bilingual vocabulary test) in three presentation formats (i.e., horizontal, vertical, and adjacent) for L2 Chinese vocabulary learning, taking L2 Chinese proficiency and test formats into consideration?

HP 8: When L2 Chinese proficiency and vocabulary test formats are taken into consideration, higher working memory capacities will be associated with higher vocabulary scores.

Results from mixed effects modeling showed that working memory capacities was a significant predictor of learning outcome as assessed by the five vocabulary test formats (C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl): generally, higher working memory capacities led to better learning outcome (i.e., higher probability of getting a correct answer). Regarding the other three vocabulary test formats (M2C_rcl, C2P_rcl, and M2P_rcl), higher working memory capacities predicted only the probability of obtaining learning gains in the recall test format of M2P_rcl. Interestingly, higher working memory capacities led to lower probability of getting a correct answer in the M2P_rcl recall task. Interactions between working memory capacities and vocabulary test formats were also found for C2M_rcg, M2C_rcg, M2P_rcg, and C2M_rcl. Specifically, for C2M_rcg, M2C_rcg, and C2M_rcl, higher working memory capacities led to better learning outcome, whereas for M2P_rcg, higher working memory capacities had a negative impact on learning outcome. Taking the results from all eight test formats, higher working memory capacities had a negative impact on M2P_rcg and M2P_rcl, which shared the pinyin element. As discussed in 5.2.1 RQ 7, higher working memory capacities supported the strategy of focusing more on the characters but less on the pinyin, which

may inevitably result in a trade-off in less learning gains on pinyin.

Results of bivariate correlations (see 4.2.6 Vocabulary Gain Scores and Working Memory Capacities) also showed a similar positive relationship between working memory capacities and learning outcome. The discrepancy in the results of bivariate correlations and mixed effects modeling lied in the three test formats of M2P_rcl, M2C_rcl, and C2P_rcl. This may be due to the difference in data levels: the correlations were performed with aggregated scores of all vocabulary formats, whereas the mixed effects modeling took into consideration the vocabulary test formats as a variable. These results partially supported the hypothesis.

The current findings that higher working memory capacities led to better learning outcome were consistent with L2 research generally (see Linck et al., 2014 for a meta-analysis), as well as studies on learning L2 English grammar (e.g., Indrarathne & Kormos, 2017) and learning L2 Chinese single characters (e.g., Kim et al., 2016). However, this study suggested that the effects of working memory capacities on learning outcome would also need to be considered together with other factors such as vocabulary test format.

5.3 Conclusion

The current study situated in computer-assisted self-study context and examined how both learner internal and external factors, namely, working memory capacities and presentation formats affected learner attention and learning outcome of L2 Chinese words. Mixed effects modeling and repeated-measures ANOVA, in addition to descriptive statistics and bivariate correlations were conducted for data analysis. Results revealed the effects of presentation formats on learning outcome and learner attention, supporting the predictions by the Cognitive Load Theory (Sweller et al., 1998) that optimized presentation formats can facilitate attention

and learning. Learner attention was also found as a significant predictor of learning outcome, but the direction of its effects can be positive or negative, depending on to which element of Chinese words learners paid attention. This finding has enriched and extended the hypothesis that paying attention to the language materials can improve L2 learning. In addition, working memory capacities was generally found as a significant predictor of learner attention and learning outcome, which has echoed previous studies on the effects of working memory capacities on learning L2 grammar and L2 vocabulary. This study has also expanded L2 research on the combined storage and executive functions of working memory by offering new evidence from L2 vocabulary learning.

As for pedagogical implications, the findings that the adjacent format was superior in promoting attention to characters and enhancing overall learning outcome than the horizontal and vertical formats would be of particular interest to teaching and learning of L2 Chinese words, especially in the classroom context. Given the novelty of the adjacent format, if teachers would like to use it for classroom teaching, preparation would need to be done in order to help and guide students to adapt to this format and develop new habits of viewing and learning that are comfortable to them. In terms of actual implementation, it may not be realistic to change all vocabulary learning materials to the adjacent format. However, a convenient starting point would be to adopt the adjacent format during class in the form of PowerPoint slides, blackboard writing, and vocabulary flashcards. Teachers may also help students in creating paper vocabulary flashcards in the adjacent format themselves. Computer programs and mobile applications can also be developed to provide electronic vocabulary flashcards in the adjacent format conveniently. Notably, it would be beneficial to explain to the students in the first place the advantage of the adjacent format, but meanwhile it is also important to make it clear that students

may just choose which presentation format they feel more comfortable in working with. That being said, teachers may allow an adapting period where students may try and learn with the adjacent format. Overall, the current findings can be applied to classroom teaching with reasonable adjustments.

APPENDICES

APPENDIX A. Target Words with Detailed Information

Table 57. Target Words with Detailed Information

No.	Word	Pinyin	English Translation	Part of Speech	Structural Configuration	Word Frequency	Stroke Number	Group
1	贫穷	pínqióng	poor	adj	TD+TD	74	15	a
2	夺取	duóqǔ	seize	v	TD+LR	70	14	a
3	忽略	hūlüè	neglect	v	TD+LR	71	19	a
4	暗示	ànshì	imply	v	LR+TD	75	18	a
5	挑选	tiāoxuǎn	choose	v	LR+HE	58	18	a
6	弥补	míbǔ	remedy	v	LR+LR	63	15	a
7	形状	xíngzhuàng	shape	n	LR+LR	70	14	a
8	脆弱	cuìruò	fragile	adj	LR+LR	76	20	a
9	循环	xúnhuán	circulate	v	LR+LR	87	20	a
10	流传	liúchuán	spread	v	LR+LR	88	16	a
11	恶劣	èliè	vile	adj	TD+TD	75	16	b
12	驾驶	jiàshǐ	drive	v	TD+LR	68	16	b
13	审判	shěnpàn	judge	v	TD+LR	87	15	b
14	欺负	qīfu	bully	v	LR+TD	85	18	b
15	强迫	qiǎngpò	compel	v	LR+HE	61	20	b
16	指挥	zhǐhuī	command	n	LR+LR	61	18	b
17	细致	xìzhì	careful	adj	LR+LR	65	18	b
18	消耗	xiāohào	consume	v	LR+LR	68	20	b
19	挖掘	wājué	excavate	v	LR+LR	72	20	b
20	珍惜	zhēnxī	treasure	v	LR+LR	78	20	b
21	姿态	zītài	posture	n	TD+TD	86	17	c
22	尖锐	jiānruì	sharp	adj	TD+LR	66	18	c
23	奖励	jiǎnglì	reward	v	TD+LR	83	16	c
24	神奇	shénqí	magical	adj	LR+TD	72	17	c

Table 57 (cont'd)

25	叙述	xùshù	narrate	v	LR+HE	68	17	c
26	讽刺	fěngcì	satirize	v	LR+LR	59	14	c
27	伴随	bànsuí	follow	v	LR+LR	65	18	c
28	谈论	tánlùn	discuss	v	LR+LR	71	16	c
29	抽烟	chōuyān	smoke	v	LR+LR	77	18	c
30	耽误	dānwù	delay	v	LR+LR	78	19	c

Notes. adj = adjective. v = verb. n = noun. LR = left-right. TD = top-down. HE = half-enclosure.
The information of pinyin, English translation, part of speech, and word frequency were collected from *A Frequency Dictionary of Mandarin Chinese* (Xiao et al., 2009), except for the following changes. The pinyin for *审判* was corrected from *shēnpàn* to *shěnpàn*, and the original English translation of *bring to trial* was replaced with *judge* (from iCIBA, an online Chinese-to-English dictionary, https://www.iciba.com/) to reduce the word length. For *叙述* and *谈论*, the original English translations of *tell about* and *talk about* were changed respectively to *narrate* and *discuss* (from iCIBA) for word length reduction. The translation of *讽刺* was changed from *satirise* to *satirize* for American English.

APPENDIX B. Radical Information of Three Word Groups

Table 58. Radical Information of Three Word Groups

Group a				
Word	New Character	Learned Character	Lesson	Radical
弥补	弥 (mí)	你 (nǐ)	1	phonetic
挑选	挑 (tiāo)	跳 (tiào)	4	
	选 (xuǎn)	先 (xiān)	1	
流传	传 (chuán)	专 (zhuān)	8*	
循环	环 (huán)	不 (bù)	1	non-phonetic
夺取	夺 (duó)	大 (dà)	3	
	取 (qǔ)	欢 (huān)	4	
	取 (qǔ)	最 (zuì)	8*	
暗示	暗 (àn)	音 (yīn)	4	
忽略	略 (lüè)	客 (kè)	4	
	略 (lüè)	路 (lù)	10*	
贫穷	贫 (pín)	分 (fēn)	6	
形状	形 (xíng)	开 (kāi)	6	
Group b				
Word	New Character	Learned Character	Lesson	Radical
欺负	欺 (qī)	期 (qī)	3	phonetic
审判	判 (pàn)	半 (bàn)	3	
消耗	耗 (hào)	毛 (máo)	9*	
驾驶	驾 (jià)	加 (jiā)	11*	
恶劣	恶 (è)	想 (xiǎng)	3	non-phonetic
	劣 (liè)	少 (shǎo)	9*	
强迫	迫 (pò)	白 (bái)	2	
	强 (qiǎng)	虽 (suī)	9*	
珍惜	惜 (xī)	错 (cuò)	4	
细致	致 (zhì)	到 (dào)	6	
挖掘	掘 (jué)	出 (chū)	10*	
Group c				
Word	New Character	Learned Character	Lesson	Radical
伴随	伴 (bàn)	半 (bàn)	3	phonetic
姿态	态 (tài)	太 (tài)	4	
	姿 (zī)	次 (cì)	13*	
谈论	论 (lùn)	伦 (lún)	14*	
抽烟	抽 (chōu)	邮 (yóu)	10*	
	烟 (yān)	因 (yīn)	3	non-phonetic
尖锐	尖 (jiān)	小 (xiǎo)	1	

Table 58 (cont'd)

	尖 (jiān)	大 (dà)	3
	锐 (ruì)	说 (shuō)	6
奖励	奖 (jiǎng)	大 (dà)	3
神奇	奇 (qí)	大 (dà)	3
	奇 (qí)	可 (kě)	3
叙述	叙 (xù)	欢 (huān)	4
	叙 (xù)	除 (chú)	8*

Note. *Lessons that were taught starting in the fifth month (i.e., second semester).

APPENDIX C. Difficulty Levels of Word Groups

Before the current, final version of the word groupings, I had developed another two versions of word groupings and had recruited eight and three participants from the target population at two universities (hereafter Pilot UK and Pilot MSU), respectively. The Pilot MSU used the same pretest and posttest as the current main study, but adopted different word groupings in statistical analyses. As for the Pilot UK, its tests differed from those of the Pilot MSU (and the current main study) in the following aspects: word groupings, two target words, availability of a recognition pretest, number of test items, time limit of tests, and format of test items (see Table 59 for details). Given these differences in the tests of the Pilot UK and the Pilot MSU (and the current main study), the estimation of the difficulty levels of the final version of the word groupings should be taken as a rational approximation at best. However, with caution in interpretation, these pilot data could be quite informative and valuable when other sources of data were not available.

Due to time constraint in data collection, I was not likely to complete analyzing the pilot data of both the production and recognition tasks before starting the current main study. Consequently, in order to finalize the instruments efficiently, I decided to focus on the pilot data of the recognition tasks, which were more readily available with the assistance of automatic scoring. Table 60 shows the facility values of the test items for the final version of the word groupings calculated with the data of Pilot MSU and Pilot UK. As mentioned previously, the tests used in Pilot MSU and Pilot UK differed in several aspects, so the facility values were calculated respectively. Facility value, also termed difficulty index, represents the percentage of test takers who provide the correct answer to the test item: the higher the facility value, the less difficult the test item (Green, 2013). As shown in the following table, in the final version of the

word groupings, each word group had similar mean facility values: .54 for Group a, .56 for Group b, and .55 for Group c in Pilot MSU; and .59, .60, and .57 for Group a, b, and c, respectively, in Pilot UK. As cautioned previously, these word difficulty levels were no more than a rough estimation to inform the groupings of target words prior to the data collection for the current main study. The Latin square design for presenting the word groups (see the following paragraph) can basically resolve the issue of unequal difficulty levels among word groups (Loewen & Plonsky, 2015; Tavakoli, 2012).

Table 59. Differences in the Tests Between Pilot MSU and Pilot UK

	Pilot MSU	Pilot UK
Target Words	细致 – careful	细致 – delicate
	抽烟 – smoke	吸烟 – smoke
Pretest	Recognition pretest	None
Item Number	Recognition: 30 items for each format (all target words covered)	Recognition: 15 items for each format (randomly chose 15 out of 30 target words)
Time	Untimed	Timed (ranged from 4s to 12s per item)
Item Format	Production (written): from meaning to characters	None
	Production (written): from meaning to pinyin	None
	Recognition (written): from characters to pinyin	Recognition: from characters (written) to pinyin (spoken)
	Recognition (written): from meaning to pinyin	Recognition: from meaning (written) to pinyin (spoken)
	Recognition (written): from pinyin to characters	None
	Recognition (written): from pinyin to meaning	None

Table 60. Item Facility Values for Pilot MSU and Pilot UK in the Final Version of Word Groupings

			Pilot MSU (3)						Pilot UK (8)				
No.	Group	Word	M2C	M2P	C2M	C2P	Mean	Word	M2C	M2P	C2M	C2P	Mean
1	a	贫穷	1.00	0.33	1.00	0.67	0.75	贫穷	0.60	0.75	0.67	0.75	0.69
2	a	夺取	0.33	0.00	0.33	0.00	0.17	夺取	0.25	0.00	0.50	0.00	0.19
3	a	忽略	0.67	0.67	1.00	0.33	0.67	忽略	0.67	0.25	0.80	0.60	0.58
4	a	暗示	0.67	0.00	1.00	0.00	0.42	暗示	1.00	1.00	0.80	0.60	0.85
5	a	挑选	0.67	0.67	0.33	1.00	0.67	挑选	0.67	0.00	0.80	0.80	0.57
6	a	弥补	1.00	0.67	1.00	1.00	0.92	弥补	0.67	0.25	0.80	0.60	0.58
7	a	形状	0.67	0.00	0.67	0.00	0.33	形状	0.67	0.25	0.80	0.25	0.49
8	a	脆弱	1.00	0.33	1.00	0.00	0.58	脆弱	0.75	0.25	0.75	0.67	0.60
9	a	循环	0.67	0.00	0.67	0.00	0.33	循环	0.75	0.25	1.00	0.33	0.58
10	a	流传	1.00	0.33	0.67	0.33	0.58	流传	1.00	0.50	1.00	0.40	0.73
Mean Facility Value			0.77	0.30	0.77	0.33	0.54		0.70	0.35	0.79	0.50	0.59
11	b	恶劣	1.00	1.00	1.00	1.00	1.00	恶劣	1.00	0.50	0.25	0.50	0.56
12	b	驾驶	1.00	0.33	1.00	0.00	0.58	驾驶	1.00	0.50	0.80	1.00	0.83
13	b	审判	1.00	0.00	1.00	0.00	0.50	审判	1.00	1.00	1.00	0.50	0.88
14	b	欺负	0.67	0.67	1.00	0.67	0.75	欺负	0.80	0.75	0.67	0.75	0.74
15	b	强迫	0.67	0.00	0.33	0.33	0.33	强迫	0.75	0.25	0.00	0.67	0.42
16	b	指挥	0.33	0.00	1.00	0.00	0.33	指挥	0.60	0.25	0.67	0.00	0.38
17	b	细致	0.50	0.00	0.50	0.00	0.25	细致	0.40	0.25	0.67	0.25	0.39
18	b	消耗	1.00	0.33	1.00	0.67	0.75	消耗	1.00	0.50	0.50	0.67	0.67
19	b	挖掘	0.67	0.00	1.00	0.00	0.42	挖掘	0.75	0.00	0.75	0.20	0.43
20	b	珍惜	1.00	0.33	1.00	0.33	0.67	珍惜	0.67	0.50	0.80	0.75	0.68
Mean Facility Value			0.78	0.27	0.88	0.30	0.56		0.80	0.45	0.61	0.53	0.60
21	c	姿态	1.00	0.67	1.00	0.33	0.75	姿态	0.75	0.67	0.50	0.75	0.67
22	c	尖锐	1.00	0.00	1.00	0.50	0.63	尖锐	1.00	0.50	0.50	0.00	0.50
23	c	奖励	0.33	0.00	0.67	0.00	0.25	奖励	0.50	0.25	0.75	1.00	0.63
24	c	神奇	1.00	0.33	1.00	0.00	0.58	神奇	1.00	0.50	0.67	0.00	0.54
25	c	叙述	0.67	0.67	0.67	0.33	0.58	叙述	0.25	0.50	0.00	0.33	0.27
26	c	讽刺	0.33	0.33	0.67	0.33	0.42	讽刺	0.60	0.00	0.67	0.75	0.50
27	c	伴随	0.67	0.00	1.00	0.33	0.50	伴随	0.80	0.25	1.00	0.67	0.68
28	c	谈论	1.00	0.67	1.00	0.67	0.83	谈论	1.00	0.75	0.33	0.50	0.65
29	c	抽烟	1.00	0.67	1.00	0.67	0.83	抽烟	1.00	1.00	1.00	0.25	0.81
30	c	耽误	0.33	0.00	0.33	0.00	0.17	耽误	0.50	0.75	0.50	0.20	0.49
Mean Facility Value			0.73	0.33	0.83	0.32	0.55		0.74	0.52	0.59	0.45	0.57

Note. M2C = from meaning to characters. M2P = from meaning to pinyin. C2M = from characters to meaning. C2P = from characters to pinyin.

APPENDIX D. Presentation Formats of Target Words in Different Word Lists

Table 61. Presentation Formats of Target Words in Different Word Lists

No.	Word	Group	List i	List ii	List iii
1	贫穷	a	H	V	A
2	夺取	a	H	V	A
3	忽略	a	H	V	A
4	暗示	a	H	V	A
5	挑选	a	H	V	A
6	弥补	a	H	V	A
7	形状	a	H	V	A
8	脆弱	a	H	V	A
9	循环	a	H	V	A
10	流传	a	H	V	A
11	恶劣	b	V	A	H
12	驾驶	b	V	A	H
13	审判	b	V	A	H
14	欺负	b	V	A	H
15	强迫	b	V	A	H
16	指挥	b	V	A	H
17	细致	b	V	A	H
18	消耗	b	V	A	H
19	挖掘	b	V	A	H
20	珍惜	b	V	A	H
21	姿态	c	A	H	V
22	尖锐	c	A	H	V
23	奖励	c	A	H	V
24	神奇	c	A	H	V
25	叙述	c	A	H	V
26	讽刺	c	A	H	V
27	伴随	c	A	H	V
28	谈论	c	A	H	V
29	抽烟	c	A	H	V
30	耽误	c	A	H	V

Note. A = Adjacent. H = Horizontal. V = Vertical.

APPENDIX E. Visual Forward Digit Span Task

Table 62. Visual Forward Digit Span Task

Trial	Length	Digits
practice	2	1 4
practice	2	7 2
practice	3	8 3 5
practice	3	2 9 6
target	3	4 7 5
target	3	3 8 6
target	4	6 1 5 8
target	4	4 2 9 7
target	5	1 8 6 2 5
target	5	4 9 5 3 7
target	6	1 7 9 3 8 4
target	6	9 1 6 2 7 5
target	7	3 9 4 8 1 5 7
target	7	4 7 1 8 2 9 3
target	8	8 3 9 6 7 4 2 1
target	8	2 7 1 3 8 6 9 5
target	9	6 8 5 9 7 2 4 1 3
target	9	5 3 6 9 2 8 4 7 1

APPENDIX F. Letter Memory Task

Table 63. Letter Memory Task

Trial	Length	Letters
practice	5	Z K Y P F
practice	5	B J Q R X
practice	7	M T D G S N C
practice	7	Q V N F L B H
target	5	Y C Z K D
target	5	P T N L J
target	5	F X S T B
target	7	V Y G K R X Q
target	7	L R H Z D Y F
target	7	T P G M S B V
target	9	C R Z F Y T S P H
target	9	K B X P N Q L G T
target	9	N S J Y V D L H C
target	11	J Z F C P D N R T G L
target	11	C F K J D Y L S P V B
target	11	G R Q M Z X T N H J C

APPENDIX G. Number Letter Task

Table 64. Number Letter Task

Number-letter Pairs							
2A	3A	4A	5A	6A	7A	8A	9A
2E	3E	4E	5E	6E	7E	8E	9E
2I	3I	4I	5I	6I	7I	8I	9I
2U	3U	4U	5U	6U	7U	8U	9U
2G	3G	4G	5G	6G	7G	8G	9G
2K	3K	4K	5K	6K	7K	8K	9K
2M	3M	4M	5M	6M	7M	8M	9M
2R	3R	4R	5R	6R	7R	8R	9R
Letter-number Pairs							
A2	A3	A4	A5	A6	A7	A8	A9
E2	E3	E4	E5	E6	E7	E8	E9
I2	I3	I4	I5	I6	I7	I8	I9
U2	U3	U4	U5	U6	U7	U8	U9
G2	G3	G4	G5	G6	G7	G8	G9
K2	K3	K4	K5	K6	K7	K8	K9
M2	M3	M4	M5	M6	M7	M8	M9
R2	R3	R4	R5	R6	R7	R8	R9

APPENDIX H. Stroop Task

Table 65. Stroop Task

Stimulus	Ink Color	Condition	Answer
Target Trials			
***	red	control	R
***	blue	control	B
***	green	control	G
***	orange	control	O
***	yellow	control	Y
***	purple	control	P
****	red	control	R
****	blue	control	B
****	green	control	G
****	orange	control	O
****	yellow	control	Y
****	purple	control	P
*****	red	control	R
*****	blue	control	B
*****	green	control	G
*****	orange	control	O
*****	yellow	control	Y
*****	purple	control	P
******	red	control	R
******	blue	control	B
******	green	control	G
******	orange	control	O
******	yellow	control	Y
******	purple	control	P
red	red	congruent	R
blue	blue	congruent	B
green	green	congruent	G
orange	orange	congruent	O
yellow	yellow	congruent	Y
purple	purple	congruent	P
red	blue	incongruent	B
red	green	incongruent	G
red	orange	incongruent	O
red	yellow	incongruent	Y

Table 65 (cont'd)

red	purple	incongruent	P
blue	red	incongruent	R
blue	green	incongruent	G
blue	orange	incongruent	O
blue	yellow	incongruent	Y
blue	purple	incongruent	P
green	red	incongruent	R
green	blue	incongruent	B
green	orange	incongruent	O
green	yellow	incongruent	Y
green	purple	incongruent	P
orange	red	incongruent	R
orange	blue	incongruent	B
orange	green	incongruent	G
orange	yellow	incongruent	Y
orange	purple	incongruent	P
yellow	red	incongruent	R
yellow	blue	incongruent	B
yellow	green	incongruent	G
yellow	orange	incongruent	O
yellow	purple	incongruent	P
purple	red	incongruent	R
purple	blue	incongruent	B
purple	green	incongruent	G
purple	yellow	incongruent	Y
purple	orange	incongruent	O
Practice Trials			
***	red	control	R
orange	yellow	incongruent	Y
****	blue	control	B
red	purple	incongruent	P
*****	green	control	G
purple	blue	incongruent	B
green	red	incongruent	R
******	orange	control	O
yellow	yellow	congruent	Y
blue	orange	incongruent	O
***	red	control	R

APPENDIX I. Test Items of Chinese Proficiency Test

Table 66. Test Items of Chinese Proficiency Test

Test Format	Test Item	Lesson	Test Source
Word Matching	水	5*	H10902
	写	7**	H10901
	菜	3*	H10901
	他	2*	H10901
Sentence Matching	他 在 睡觉 呢。	8***	Sample
	来，我们 看看 里面 是 什么 东西。	13***	H10901
	不客气，王先生，请 坐。	6**	H11003
	他们 是 同学。	3*	Sample
Sentence Selection[1]	那 几 本 书 怎么样？ - 都 很 好。	4*	H10902
	他 女儿 多 大 了？ - 7 岁。	3*	H10901
	医生！医生 在 哪儿？ - 那儿。	8***	H10902
	爸爸 什么 时候 来 北京 呢？ - 下 个 月。	6**	H10901
Word Selection[1]	昨天 是 8 月 19 日。	3*	H10901
	男：北京 昨天 天气 怎么样？ 女：很 冷。	11***	H11003
	喂，张先生 在 家 吗？	6**	Sample
	男：你 认识 他？他 是 谁？ 女：他 是 我 的 学生。	2*	H10902

Note. *Lesson 1-5. **Lesson 6-7. ***Lesson 8-13.

[1] For Sentence Selection and Word Selection, the fill-in-the-blank items are presented here with answer keys (i.e., the underlined words).

APPENDIX J. Interview Questions

1. Let's look at your ratings. Here you gave horizontal format ____ stars, vertical format ____ stars, and adjacent format ____ stars. Could you explain why?
2. When you were studying the Chinese words just now over there,
    a. how did you learn the characters, the pinyin, and the meaning in each format?
    b. In the horizontal format, what did you start with, and what was the next?
    c. What about the vertical format? The adjacent format?
3. When you were studying the Chinese words just now over there, did you use any strategies?
    a. Could you give me an example?
    b. Did you use the same strategies for each format?

APPENDIX K. Missing Data in the Working Memory Tasks

The field of missing data research has long been arguing for more sophisticated treatment to missing values rather than simply deleting the entire case with missing values (e.g., Enders, 2010; Little & Rubin, 2020; Raghunathan, 2015). According to Little and Rubin (2020), "missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value" (p. 4). Similarly, Raghunathan's (2015) definition of missing data highlights the meaningful value that is hidden. Although the rate of missing data affects which particular remedy to use, the missing data mechanism is more important in deciding the appropriate methods to avoid biased results (Enders, 2010; Little & Rubin, 2020; Raghunathan, 2015). In multivariate analysis, based on the relationship from the variable with missing values to other variables and to itself, there are three major missing data mechanisms (Enders, 2010; Little & Rubin, 2020; Raghunathan, 2015). If the missing value of a variable is not related to the values of other variables or itself, the mechanism is missing completely at random (MCAR). If it is related to the values of other variables rather than itself, the mechanism is missing at random (MAR). Lastly, if it is related to itself rather than the values of other variables, the mechanism is missing not at random (MNAR).

According to the definition of missing data mechanisms, the missing accuracy rate in the letter memory task was MCAR, because the participant discontinued the task due to personal reason. For the number letter task, 1 participant gave up the task because of personal reason and led to MCAR. Another 8 participants' RT differences were regarded as invalid due to their low accuracy rates in the task. These discarded RT differences belonged to MAR, because they were missed due to the values of another variable, the accuracy rate, but not that they were beyond a certain RT difference range (the values themselves). For the Stroop task, 2 participants self-

identified as color-blind, so their RT differences were MCAR. Another 2 participants' RT differences were regarded as invalid due to their low accuracy rates in this task and led to MAR. In summary, the missing data mechanisms were MCAR or MAR for the working memory tasks.

Common methods to deal with missing data include skipping the entire case with any missing value in all analyses, even if some of the values are available for some analyses (i.e., listwise deletion). Another common way is to exclude the case from the analysis only when its value for that particular analysis is missing, and run other analyses with other available values (i.e., pairwise deletion). However, both deletion methods are found to generate biased results when the missing data mechanism is not MCAR (Enders, 2010; Little & Rubin, 2020; Raghunathan, 2015). In addition, listwise deletion suffers the disadvantage of reducing the sample size, lowering the statistical power, and wasting the enormous efforts in collecting data (Hair et al., 2018). Pairwise deletion often leads to unequal numbers of cases in different analyses and can generate statistical values that are beyond the possible range (Hair et al., 2018). The missing data literature has advocated more sophisticated methods to impute the missing values (Enders, 2010; Little & Rubin, 2020; Raghunathan, 2015). Although multiple imputation has been widely recommended as an effective way to deal with missing data (Enders, 2010; Little & Rubin, 2020; Raghunathan, 2015; van Buuren, 2018), for principal component analysis, multiple imputation methods are still developing and the difficulty lies in the important step of pooling the results of multiple imputed data sets (van Ginkel, Linting, Rippe, & van der Voort, 2019; van Ginkel & Kroonenberg, 2014). In fact, on the CRAN Task View page dedicated to missing data (Josse, Tierney, & Vialaneix, 2020), most of the R packages for principle component analysis adopt single imputation principles, including the missMDA (Josse & Husson, 2016) package.

The missMDA package uses a regularized iterative expectation-maximization algorithm (EM-PCA) to complete the data set with missing values (continuous, categorical, and mixed data), which can then be used for principal component analysis and any other statistical analysis (Josse & Husson, 2016). Since EM-based methods generally accommodate both MCAR and MAR (Hair et al., 2018), the missMDA is appropriate for the working memory data in this study. In addition, compared with other R packages for principle component analysis with missing data, the missMDA has many companion materials, including several Youtube tutorials. Therefore, I used RStudio 1.1.447 with missMDA to impute the missing values in the working memory tasks.

APPENDIX L. Mixed Effects Models for RQ 3

Table 67. #1. Fixation Durations of Characters (1): Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl.

Formula					
Gain ~ Prof + TF + Dur.C + Dur.C:TF + (1 + PF + Dur.C | ID) + (1 + PF | Word) + (1 + Dur.C | Word)					
Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.568*	0.175	-3.251	.001
Prof		0.704*	0.258	2.733	.006
M2C_rcg		0.864*	0.048	18.059	< .001
M2P_rcg		-0.222*	0.048	-4.656	< .001
C2M_rcl		-1.090*	0.056	-19.375	< .001
C2M_rcg		1.025*	0.049	21.054	< .001
Dur.C		0.678*	0.133	5.087	< .001
M2C_rcg:Dur.C		0.250*	0.102	2.455	.014
M2P_rcg:Dur.C		-1.077*	0.100	-10.757	< .001
C2M_rcl:Dur.C		0.769*	0.128	5.990	< .001
C2M_rcg:Dur.C		0.477*	0.104	4.587	< .001
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	1.020	1.010		
	PF-A	0.114	0.338		
	PF-H	0.069	0.262		
	Dur.C	0.635	0.797		
Word	Intercept	0.226	0.476		
	PF-A	0.028	0.168		
	PF-H	0.034	0.185		
Word	Intercept	0.215	0.464		
	Dur.C	0.070	0.264		

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = Proficiency. Dur.C = Fixation Duration of Characters.

Table 68. #2. Fixation Durations of Characters (2): Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl.

		Continuous				Binary			
Formula		Gain ~ 1 + (1 + Dur.C | ID) + (1 + Dur.C | Word)				Gain ~ Prof + TF + Dur.C + Dur.C:TF + (0 + Dur.C | ID) + (0 + Dur.C | Word)			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.570*	0.097	-5.855	< .001	-4.497*	0.167	-26.948	< .001
Prof						0.939*	0.228	4.120	< .001
M2C_rcl						-1.389*	0.272	-5.100	< .001
M2P_rcl						-0.405*	0.200	-2.022	.043
Dur.C						0.415	0.377	1.100	.272
M2C_rcl:Dur.C						0.681	0.522	1.309	.191
M2P_rcl:Dur.C						-0.787*	0.368	-2.138	.033
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.039	0.198						
	Dur.C	0.131	0.363			1.623	1.274		
Word	Intercept	0.029	0.170						
	Dur.C	0.085	0.291			0.718	0.848		

Note. *$p < .05$.

Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Prof = L2 Chinese Proficiency. Dur.C = Fixation Duration of Characters.

Table 69. #3. Fixation Durations of Pinyin (1): Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl.

Formula					
Gain ~ Prof + TF + Dur.P + Dur.P:TF + (1 + PF + Dur.P | ID) + (1 + PF | Word) + (1 + Dur.P | Word)					
Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.227	0.194	-1.171	.241
Prof		0. 627*	0.263	2. 379	.017
M2C_rcg		1.081*	0.095	11.353	< .001
M2P_rcg		-0.942*	0.094	-9.995	< .001
C2M_rcl		-0.738*	0.105	-7.036	< .001
C2M_rcg		1.369*	0.097	14.074	< .001
Dur.P (rescaled)		-1.567 *	0. 393	-3.984	< .001
M2C_rcg:Dur.P (rescaled)		-0.993*	0.352	-2.822	.005
M2P_rcg:Dur.P (rescaled)		2.973*	0.346	8.584	< .001
C2M_rcl:Dur.P (rescaled)		-1.192*	0.424	-2.808	.005
C2M_rcg:Dur.P (rescaled)		-1.541*	0.356	-4.324	< .001
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.978	0.989		
	PF-A	0.091	0.302		
	PF-H	0.072	0.268		
	Dur.P (rescaled)	4.388	2.095		
Word	Intercept	0.219	0.468		
	PF-A	0.033	0.183		
	PF-H	0.038	0.193		
Word	Intercept	0.336	0.580		
	Dur.C	0.765	0.875		

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = Proficiency. Dur.P = Fixation Duration of Pinyin.

Table 70. #4. Fixation Durations of Pinyin (2): Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl.

		Continuous				Binary			
Formula		Gain ~ 1 + (0 + Dur.P | ID) + (1 | Word)				Gain ~ Prof + TF + Dur.P + Dur.P:TF + (1 + Dur.P | ID) + (0 + Dur.C | Word)			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.488*	0.062	-7.810	< .001	-5.033*	0.247	-20.399	< .001
Prof						0.750*	0.364	2.062	.039
M2C_rcl						-1.378*	0.270	-5.113	< .001
M2P_rcl						-0.504*	0.208	-2.427	.015
Dur.P						-0.213	0.449	-0.475	.635
M2C_rcl:Dur.P						-0.813	0.573	-1.419	.156
M2P_rcl:Dur.P						0.889*	0.377	2.359	.018
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept					1.384	1.177		
	Dur.P	0.030	0.172			0.095	0.308		
Word	Intercept	0.017	0.131						
	Dur.P					0.835	0.914		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Prof = L2 Chinese Proficiency. Dur.P = Fixation Duration of Pinyin.

Table 71. #5. Fixation Durations of Meaning (1): Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl.

Formula					
Gain ~ PF + Prof + TF + Dur.M + Dur.M:TF + (1 + PF + Dur.M: | ID) + (1 + PF + Dur.M | Word)					
Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.611*	0.168	-3.645	< .001
PF-A		0.129*	0.056	2.301	.021
PF-H		-0.075	0.057	-1.325	.185
Prof		0.778*	0.236	3.301	.001
M2C_rcg		0.846*	0.047	17.963	< .001
M2P_rcg		-0.232*	0.047	-4.904	< .001
C2M_rcl		-1.024*	0.054	-19.077	< .001
C2M_rcg		1.001*	0.048	20.936	< .001
Dur.M		-0.490*	0.120	-4.101	< .001
M2C_rcg:Dur.M		-0.041	0.100	-0.411	.681
M2P_rcg:Dur.C		0.592*	0.100	5.915	< .001
C2M_rcl:Dur.C		-0.415*	0.126	-3.303	.001
C2M_rcg:Dur.C		-0.260*	0.101	-2.570	.010
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.947	0.973		
	PF-A	0.080	0.282		
	PF-H	0.074	0.272		
	Dur.M	0.375	0.612		
Word	Intercept	0.405	0.636		
	PF-A	0.021	0.144		
	PF-H	0.026	0.160		
	Dur.M	0.033	0.181		

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = Proficiency. Dur.M = Fixation Duration of Meaning.

Table 72. #6. Fixation Durations of Meaning (2): Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl.

		Continuous				Binary			
Formula		Gain ~ 1 + (1 + Dur.M | ID) + (1 + Dur.M | Word)				Gain ~ Prof + TF + Dur.M + Dur.M:TF + (1 + Dur.M | ID) + (1 | Word)			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.655*	0.099	-6.587	< .001	-5.597*	0.346	-16.178	< .001
Prof						0.702	0.389	1.805	.071
M2C_rcl						-1.405*	0.244	-5.757	< .001
M2P_rcl						-0.553*	0.196	-2.825	.005
Dur.M						-1.232*	0.487	-2.532	.011
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.036	0.190			1.761	1.327		
	Dur.M	0.083	0.287			2.171	1.473		
Word	Intercept	0.031	0.177			0.898	0.948		
	Dur.M	0.197	0.444						

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Prof = L2 Chinese Proficiency. Dur.M = Fixation Duration of Meaning.

Table 73. #7. Fixation Counts of Characters (1): Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl.

Formula					
Gain ~ Prof + TF + Count.C + Count.C:TF + (1 + PF + Count.C | ID) + (1 + PF + Count.C | Word)					
Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.534*	0.174	-3.063	.002
Prof		0.539*	0.259	2.081	.037
M2C_rcg		0.853*	0.048	17.803	< .001
M2P_rcg		-0.223*	0.048	-4.695	< .001
C2M_rcl		-1.060*	0.055	-19.324	< .001
C2M_rcg		1.017*	0.049	20.831	< .001
Count.C (rescaled)		0.678*	0.132	5.138	< .001
M2C_rcg:Count.C (rescaled)		0.399*	0.101	3.943	< .001
M2P_rcg:Count.C (rescaled)		-1.083*	0.100	-10.882	< .001
C2M_rcl:Count.C (rescaled)		0.463*	0.116	3.986	< .001
C2M_rcg:Count.C (rescaled)		0.611*	0.104	5.879	< .001
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.983	0.992		
	PF-A	0.120	0.347		
	PF-H	0.055	0.235		
	Count.C (rescaled)	0.741	0.861		
Word	Intercept	0.439	0.662		
	PF-A	0.035	0.187		
	PF-H	0.033	0.181		
	Count.C (rescaled)	0.091	0.302		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = Proficiency. Count.C = Fixation Count of Characters.

Table 74. #8. Fixation Counts of Characters (2): Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl.

		Continuous				Binary			
Formula		Gain ~ 1 + (1 + Count.C | ID) + (1 | Word)				Gain ~ Prof + TF + Count.C + Count.C:TF + (1 + Count.C | ID) + (0 + Count.C | Word)			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.555*	0.086	-6.422	< .001	-5.059*	0.255	-19.866	< .001
Prof						0.623	0.371	1.681	.093
M2C_rcl						-1.479*	0.289	-5.110	< .001
M2P_rcl						-0.424*	0.210	-2.023	.043
Count.C						0.580	0.431	1.345	.179
M2C_rcl:Count.C						0.846	0.513	1.651	.099
M2P_rcl:Count.C						-0.836*	0.387	-2.161	.031
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.036	0.189			1.407	1.186		
	Count.C	0.229	0.479			0.706	0.840		
Word	Intercept	0.022	0.149						
	Count.C					0.807	0.898		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Prof = L2 Chinese Proficiency. Count.C = Fixation Count of Characters.

Table 75. #9. Fixation Counts of Pinyin (1): Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl.

Formula					
Gain ~ PF + Prof + TF + Count.P + Count.P:PF + Count.P:TF + (1 + PF + Count.P | ID) + (1 + PF + Count.P | Word)					
Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.617*	0.172	-3.578	< .001
PF-A		0.041	0.068	0.600	.548
PF-H		-0.029	0.059	-0.495	.621
Prof		0.781*	0.258	3.026	.002
M2C_rcg		0.847*	0.047	17.971	< .001
M2P_rcg		-0.242*	0.047	-5.106	< .001
C2M_rcl		-1.018*	0.053	-19.149	< .001
C2M_rcg		1.004*	0.048	21.034	< .001
Count.P		0.286*	0.132	-2.163	.031
PF-A:Count.P		-0.042	0.101	-0.415	.678
PF-H:Count.P		-0.202*	0.091	-2.222	.026
M2C_rcg:Count.P		0.225*	0.097	-2.315	.021
M2P_rcg:Count.P		0.614*	0.097	6.334	< .001
C2M_rcl:Count.P		-0.436*	0.115	-3.779	< .001
C2M_rcg:Count.P		-0.243*	0.098	-2.489	.013
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	1.037	1.018		
	PF-A	0.083	0.288		
	PF-H	0.056	0.237		
	Count.P	0.366	0.605		
Word	Intercept	0.406	0.637		
	PF-A	0.050	0.223		
	PF-H	0.037	0.193		
	Count.P	0.130	0.360		

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = Proficiency. Count.P = Fixation Count of Pinyin.

Table 76. #10. Fixation Counts of Pinyin (2): Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl.

		Continuous				Binary			
Formula		Gain ~ 1 + (1 + Count.P | ID) + (1 + Count.P | Word)				Gain ~ Prof + TF + (0 + Count.C | ID) + (0 + Count.C | Word)			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.608*	0.094	-6.463	< .001	-4.457*	0.152	-29.273	< .001
Prof						0.971*	0.221	4.403	< .001
M2C_rcl						-1.284*	0.234	-5.499	< .001
M2P_rcl						-0.472*	0.187	-2.529	.011
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.042	0.204						
	Count.P	0.068	0.260			1.265	1.125		
Word	Intercept	0.025	0.157						
	Count.P	0.041	0.203			0.859	0.927		

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Prof = L2 Chinese Proficiency. Count.P = Fixation Count of Pinyin.

Table 77. #11. Fixation Counts of Meaning (1): Mixed Logit Model for C2M_rcg, M2C_rcg, C2P_rcg, M2P_rcg, and C2M_rcl.

Formula					
Gain ~ PF + Prof + TF + Count.M + Count.M:TF + (1 + PF + Count.M | ID) + (1 + PF | Word) + (1 + Count.M | Word)					
Fixed Effects					
		Estimate	*Std.Error*	*z*	*p*
Intercept		-0.271	0.182	-1.493	.135
PF-A		0.121*	0.056	2.167	.030
PF-H		-0.089	0.055	-1.612	.107
Prof		0. 801*	0.239	3.356	.001
M2C_rcg		0.810*	0.097	8.380	< .001
M2P_rcg		-0.557*	0.097	-5.725	< .001
C2M_rcl		-0.758*	0.108	-7.028	< .001
C2M_rcg		1.121*	0.098	11.410	< .001
Count.M (rescaled)		-1.690 *	0.527	-3.209	< .001
M2C_rcg:Count.M (rescaled)		0.179*	0.443	0.405	.686
M2P_rcg:Count.M (rescaled)		1.622*	0.449	3.616	< .001
C2M_rcl:Count.M (rescaled)		-1.275*	0.527	-2.418	.016
C2M_rcg:Count.M (rescaled)		-0.624*	0.446	-1.398	.162
Random Effects					
		Variance	*Std.Dev.*		
ID	Intercept	0.966	0.983		
	PF-A	0.074	0.273		
	PF-H	0.061	0.246		
	Count.M (rescaled)	7.451	2.730		
Word	Intercept	0.302	0.550		
	PF-A	0.023	0.150		
	PF-H	0.027	0.165		
Word	Intercept	0.115	0.339		
	Count.M (rescaled)	0.949	0.974		

Note. $^{*}p < .05$.
Std.Error = Standard Error. *Std.Dev*. = Standard Deviation. PF-A = Adjacent. PF-H = Horizontal. Prof = Proficiency. Count.M = Fixation Count of Meaning.

Table 78. #12. Fixation Counts of Meaning (2): Two-Part Mixed Effects Model for M2C_rcl, C2P_rcl, and M2P_rcl

		Continuous				Binary			
Formula		Gain ~ 1 + (1 + Count.M | ID) + (1 + Count.M | Word)				Gain ~ Prof + TF + Count.M + (1 + Count.C | ID) + (1 | Word)			
		Fixed Effects							
		Estimate	*Std.Error*	*z*	*p*	*Estimate*	*Std.Error*	*z*	*p*
Intercept		-0.647*	0.097	-6.708	< .001	-5.542*	0.335	-16.554	< .001
Prof						0.710	0.389	1.825	.068
M2C_rcl						-1.405*	0.244	-5.758	< .001
M2P_rcl						-0.553*	0.196	-2.825	.005
Count.M						-0.890*	0.415	-2.143	.032
		Random Effects							
		Variance	*Std.Dev.*			*Variance*	*Std.Dev.*		
ID	Intercept	0.034	0.185			1.727	1.314		
	Count.M	0.007	0.085			1.796	1.340		
Word	Intercept	0.029	0.172			0.893	0.945		
	Count.M	0.135	0.367						

Note. *$p < .05$.
Std.Error = Standard Error. *Std.Dev.* = Standard Deviation. Prof = L2 Chinese Proficiency. Count.M = Fixation Count of Meaning.

REFERENCES

REFERENCES

Allen, I. E., & Seaman, J. (2010). *Class differences: Online education in the United States, 2010*. The Sloan Consortium.

Allen, I. E., Seaman, J., Poulin, R., & Straut, T. T. (2016). *Online report card: Tracking online education in the United States*. Retrieved from https://onlinelearningconsortium.org/read/online-report-card-tracking-online-education-united-states-2015/

Al-Shehri, S., & Gitsaki, C. (2010). Online reading: A preliminary study of the impact of integrated and split-attention formats on L2 students' cognitive load. *ReCALL*, *22*(3), 356-375.

Atkins, P. W. B., & Baddeley, A. D. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics*, *19*, 537-552.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. New York: Oxford University Press.

Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. D. (2007). *Working memory, thought and action*. Oxford, UK: Oxford University Press.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*(1), 1–29.

Baddeley, A. D. (2015). Working memory in second language learning. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 17-28). Bristol: Multilingual Matters.

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158-173.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47-89). New York: Academic Press.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158-175.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology*, *4*, 328.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General*, *132*(1), 71-92.

Bell, P. (2009). Le cadeau or la cadeau? The role of aptitude in learner awareness of gender distinctions in French. *The Canadian Modern Language Review*, *65*(4), 615-643.

Benati, A. (2016). Input manipulation, enhancement and processing: Theoretical views and empirical research. *Studies in Second Language Learning and Teaching*, *6*(1), 65-88.

Bojarskaite, L., Bjørnstad, D. M., Pettersen, K. H., Cunen, C., Hermansen, G. H., Åbjørsbråten, K. S., Chambers, A. R., Sprengel, R., Vervaeke, K., Tang, W., & Enger, R. (2020). Astrocytic Ca 2+ signaling is reduced during sleep and is involved in the regulation of slow wave sleep. *Nature Communications*, *11*(1), 1-16.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A, Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., Bolker, B. M. (2017). glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Bunting, M. F., & Engle, R. W. (2015). Foreword. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. xvii-xxiv). Bristol: Multilingual Matters.

Chen, X. (2013). *Chinese EFL learners' noticing of recasts: Its relation to target structure, uptake, and working memory capacity* (Doctoral dissertation). East Lansing, Michigan: Michigan State University.

Chinese Testing International. (n.d.). *HSK Level I*. Retrieved from http://www.chinesetest.cn/gosign.do?id=1&lid=0#

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73-101.

Chun, D., Smith, B., & Kern, R. (2016). Technology in language use, language teaching, and language learning. *The Modern Language Journal*, *100*, 64-80.

Chung, K. K. H. (2007). Presentation factors in the learning of Chinese characters: The order and position of Hanyu pinyin and English translations. *Educational Psychology*, *27*(1), 1-20.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

Clinefelter, D. L., & Aslanian, C. B. (2016). *Online college students 2016: Comprehensive data on demands and preferences*. Louisville, KY: The Learning House, Inc.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, *34*(1), 158-171.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769-786.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics*, *5*, 161–170.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42-45.

Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: A comment on Baguley (2012). *Behavior Research Methods*, *46*(4), 1149-1151.

Cowan, N. (1995). *Attention and memory: An integrated framework*. New York: Oxford University Press.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). Cambridge, UK: Cambridge University Press.

Cowan, N. (2005). *Working memory capacity*. Hove, East Sussex, UK: Psychology Press.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In W. Sossin, J.-C. Lacaille, F. V. Castellucci, & S. Belleville (Eds.), *Progress in brain research: The essence of memory* (pp. 323-338). Amsterdam: Elsevier/Academic Press.

Cowan, N. (2015). Second language use, theories of working memory and the Vennian mind. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 29-40). Bristol: Multilingual Matters.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158-1170.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*(3), 369-382.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), Advancing quantitative methods in second language research (pp.159-181). New York, NY: Routledge.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Dixon, M. L., Fox, K. C. R., & Christoff, K. (2014). A framework for understanding the relationship between externally and internally directed cognition. *Neuropsychologia*, *62*, 321-330.

Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics*, *30*, 149-166.

Elsabry, E., & Sumikura, K. (2020). Does open access to academic research help small, science-based companies? *Journal of Industry-University Collaboration*, 1-15.

Enders, C. K. (2010). *Applied missing data analysis*. New York: The Guilford Press.

Engle, R. W. (2001). What is working memory capacity? In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297-314). Washington, DC: American Psychological Association.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 145-199). New York, NY: Academic Press.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309-331.

Everson, M. E. (2011). Best practices in teaching logographic and non-Roman writing systems to L2 learners. *Annual Review of Applied Linguistics*, *31*, 249-274.

Farewell, V. T., Long, D. L., Tom, B. D. M., Yiu, S., & Su, L. (2017). Two-part and related regression models for longitudinal data. *Annual Review of Statistics and Its Application*, *4*, 283-315.

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th Ed.). Los Angeles: SAGE.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*, 226-236.

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: Generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic Bulletin & Review*, *19*(3), 395-404.

Frenck-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings. *Second Language Research*, *21*(2), 175-198.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*, 201-225.

Friedman, N. P., Miyake, A., Altamirano, L. J., Corley, R. P., Young, S. E., Rhea, S. A., & Hewitt, J. K. (2016). Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. *Developmental Psychology*, *52*(2), 326-340.

Gass, S. M. (2010). The relationship between L2 input and L2 output. In E. Macaro (Ed.), *The Bloomsbury companion to second language acquisition* (pp. 194-219). London, GBR: Continuum International Publishing.

Gass, S. M., & Lee, J. (2011). Working memory capacity, inhibitory control, and proficiency in a second language. In M. S. Schmid & W. Lowie (Eds.), *Modeling bilingualism: From structure to chaos: In honor of Kees De Bot* (pp. 59-84). Amsterdam, NLD: John Benjamins Publishing Company.

Gass, S. M., & Mackey, A. (2015). Input, interaction and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 180-206). New York: Routledge.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*, 513-543.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel-hierarchical models*. Cambridge: Cambridge University Press.

Genc, H., & Gülözer, K. (2013). The effect of cognitive load associated with instructional formats and types of presentation on second language reading comprehension performance. *The Turkish Online Journal of Educational Technology*, *12*(4), 171-182.

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction*, *16*, 511-525.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words. Gauging the role of attention in Incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*, 483-517.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports, and vocabulary learning scores. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 183-205). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning*, *65*(4), 896-928.

Godfroid, A., & Uggen, M. S. (2013). Attention to irregular verbs by beginning learners of German. An eye-movement study. *Studies in Second Language Acquisition*, *35*, 291-322.

Godfroid, A., Winke, P., & Conklin, K. (2020). Exploring the depths of second language processing with eye tracking: An introduction. *Second Language Research*, *36*(3), 243-255.

Goldstein, H., & Healy, M. J. R. (1995). MJR: The graphical presentation of a collection of means. *Journal of the Royal Statistical Society: Series A*, *158*, 175–177.

Graesser A. C., Halpern D. F., & Hakel M. (2008). *25 principles of learning*. Washington, DC: Task Force on Lifelong Learning at Work and at Home.

Green, R. (2013). *Statistical analyses for language testers*. Basingstoke: Palgrave Macmillan.

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*(2), 165-198.

Guan, C. Q., Liu, Y., Chan, D. H. L., Ye, F., & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology*, *103*(3), 509-522.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Hampshire: Cengage Learning EMEA.

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, *18*, 1279-1296.

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. *Applied Linguistics*, *29*, 597-618.

Hardin, J. W., & Hilbe, J. M. (2018) Generalized linear models and extensions. College Station, TX: Stata Press.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, *14*, 25-38.

Hayes, A. F. (2016). *PROCESS for SPSS 2.16*. Retrieved from http://www.processmacro.org/download.html

Hayes, A. F., & Montoya, A. K. (2017). A tutorial on testing, visualizing, and probing an interaction involving a multicategorical variable in linear regression analysis. *Communication Methods and Measures*, *11*(1), 1-30.

Hayes, A. F., & Preacher, K. J. (2013). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*, 451-470.

Heift, T., & Chapelle, C. A. (2012). Language learning through technology. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 555-569). New York: Routledge.

Henderson, J. M., & Luke, S. G. (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1390-1400.

Hernández, T. (2011). Re-examining the role of explicit instruction and input flood on the acquisition of Spanish discourse markers. *Language Teaching Research*, *15*, 159-182.

Hohenstein, S., Matuschek, H., & Kliegl, R. (2017). Linked linear mixed models: A joint analysis of fixation locations and fixation durations in natural reading. *Psychonomic bulletin & review*, *24*(3), 637-651.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology*, *23*, 1215-1226.

Hsiao, Y. P., & Broeder, P. (2013). Applying the writing scales of the Common European Framework of Reference for Languages to the new HSK test of proficiency in Chinese: Realities, problems and some suggestions for Chinese language teachers and learners. *Language Learning in Higher Education*, *2*(1), 59-74.

Hu, B. (2010). The challenges of Chinese: A preliminary study of UK learners' perceptions of difficulty. *The Language Learning Journal*, *38*(1), 99-118.

Hung, H. C. M. (2007). Reducing extraneous cognitive load by using integrated format in reading comprehension for EFL/ESL. In C. Gitsaki (Ed.), *Language and languages: Global and local tensions* (pp. 130-146). Newcastle, UK: Cambridge Scholars Publishing.

Hung, H.-T. (2015). Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning*, *28*(1), 81-96.

Hunt, D., Stuart, S., Nell, J., Hausdorff, J. M., Galna, B., Rochester, L., & Alcock, L. (2018). Do people with Parkinson's disease look at task relevant stimuli when walking? An exploration of eye movements. *Behavioural Brain Research*, *348*, 82-89.

Indrarathne, B., & Kormos, J. (2016). Attentional processing of input in explicit and implicit conditions. An eye-tracking study. *Studies in Second Language Acquisition*, 1-30.

Indrarathne, B., & Kormos, J. (2017). The role of working memory in processing L2 input: Insights from eye-tracking. *Bilingualism: Language and Cognition*, 1-20.

Issa, B., & Morgan-short, K. (2019). Effects of external attentional and internal manipulations on second language grammar development. An eye-tracking study. *Studies in Second Language Acquisition*, *41*, 389-417.

Issa, B., Morgan-Short, K., Villegas, B., & Raney, G. (2015). An eye-tracking study on the role of attention and its relationship with motivation. In L. Roberts, K. McManus, N. Vanek, & D. Trenkic (Eds.), *EUROSLA Yearbook 2015* (pp. 114-142). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Jackson, F. H., & Malone, M. E. (2010). *Building the foreign language capacity we need: Toward a comprehensive strategy for a national language framework*. Washington, DC: Center of Applied Linguistics.

Jaeger, (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446.

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied*, *18*(2), 178-191.

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, *70*(1), 1-31.

Josse, J., Tierney, N., & Vialaneix, N. (2020). *CRAN task view: Missing data*. Retrieved from https://cran.r-project.org/web/views/MissingData.html

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(2), 137-166.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122-149.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika*, *35*, 401–415.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark 4. *Educational and Psychological Measurement*, *34*(1), 111–117.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*, 23-32.

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors*, *40*, 1-17.

Kalyuga, S., & Renkl, A. (2010). Expertise reveal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, *38*, 209-215.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189-217.

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21-48). New York: Oxford University Press.

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2011). The relationship between vocabulary and short-term memory measures in monolingual and bilingual speakers. *International Journal of Bilingualism*, *15*(4), 408-425.

Ke, C. (1996). An empirical study on the relationship between Chinese character recognition and production. *The Modern Language Journal*, *80*(3), 340-350.

Ke, C., Wen, X., & Kotenbeutel, C. (2001). Report on the 2000 CLTA articulation. *Journal of the Chinese Language Teachers Association*, *36*, 23-58.

Kim, S.-A., Christianson, K., & Packard, J. (2015). Working memory in L2 character processing: The case of learning to read Chinese. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 85-104). Bristol: Multilingual Matters.

Kim, S. A., Packard, J., Christianson, K., Anderson, R. C., & Shin, J. A. (2016). Orthographic consistency and individual learner differences in second language literacy acquisition. *Reading and Writing*, *29*(7), 1409-1434.

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory: L2 question development through recasts in a laboratory setting. *Studies in Second Language Acquisition*, *37*, 549-581.

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*, 228-242.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*, 261-271.

LaFlair, G., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46–77). New York, NY: Routledge.

Lai, C., Fei, F., & Roots, R. (2008). The contingency of recasts and noticing. *CALICO Journal*, *26*(1), 70-90.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. London, England: Routledge.

Larson-Hall, J., & Plonsky, L. (2015). Chapter 6 Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(Suppl. 1), 127-159.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202-226.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436.

Lee, C. H., & Kalyuga, S. (2011). Effectiveness of different pinyin presentation formats in learning Chinese characters: A Cognitive Load perspective. *Language Learning*, *61*(4), 1099-1118.

Lee, S. K., & Huang, H. T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, *30*, 307-331.

Leow, R. P. (1999a). Attention, awareness, and Focus on Form research: A critical overview. In J. F. Lee & A. Valdman (Eds.), *Form and meaning: Multiple perspectives* (pp. 69-96). Boston, MA: Heinle & Heinle.

Leow, R. P. (1999b). The role of attention in second/foreign language classroom research: Methodological issues. In F. Martinez-Gil & J. Gutierrez-Rexac (Eds.), *Advances in Hispanic linguistics: Papers from the 2nd Hispanic Linguistics Symposium* (pp. 60-71). Somerville, MA: Cascadilla Press.

Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research*, *30*(2), 111-127.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal*, *97*(3), 634-654.

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, *58*, 259-289.

Lin, Y. Y., Holmqvist, K., Miyoshi, K., & Ashida, H. (2017). Effects of detailed illustrations on science learning: An eye-tracking study. Instructional Science, *45*(5), 557-581.

Linck, J. A., & Cunnings, I. (2015). Chapter 8 The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(Suppl. 1), 185-207.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*(3), 530-566.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*, 861-883.

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Hoboken: John Wiley 7 Sons, Inc.

Liu, L., Shih, Y.-C. T., Strawderman, R. L., Zhang, D., Johnson, B. A., & Chai, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data: A review. *Statistical Science*, *34*(2), 253-279.

Liu, Y., Wang, M., & Perfetti, C. A. (2007). Threshold-style processing of Chinese characters for adult second-language learners. *Memory & Cognition*, *35*(3), 471-480.

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An In Vivo experiment. *Language Learning*, *61*(4), 1119-1141.

Liu, Y., Yao, T., Bi, N.-P., Ge, L., & Shi, Y. (2016). *Integrated Chinese Level 1 Volume 1* (4th ed.). Boston, MA: Cheng & Tsui.

Liu, Y., Yao, T., Bi, N.-P., Ge, L., & Shi, Y. (2017). *Integrated Chinese Leve 1 Volume 2* (4th ed.). Boston, MA: Cheng & Tsui.

Loewen, S., & Inceoglu, S. (2016). The effectiveness of visual input enhancement on the noticing and L2 development of the Spanish past tense. *Studies in Second Language Learning and Teaching*, *6*(1), 89-110.

Loewen, S. & Plonsky, L. (2015). *An A-Z of applied linguistics research methods*. New York: Palgrave Macmillan.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476-490.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). Amsterdam: John Benjamins.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego: Academic Press.

Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2020). performance: Assessment of regression models performance. *CRAN*.

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 & P. Skehan (Eds.), *Individual differences and instructed language learning* (pp. 181-209). Philadelphia, NL: John Benjamins Publishing Company.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163-203.

Man, K., & Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement*, *79*(4), 617-635.

Man, K., & Harring, J. R. (2020). Assessing Preknowledge Cheating via Innovative Measures: A Multiple-Group Analysis of Jointly Modeling Item Responses, Response Times, and Visual Fixation Counts. *Educational and Psychological Measurement*.

Marefat, H., Rezaee, A. A., & Naserieh, F. (2016). Effect of computerized gloss presentation format on reading comprehension: A Cognitive Load perspective. *Journal of Information Technology Education: Research*, *15*, 479-501.

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*, 379-413.

Mason, L., Pluchino, P., Tornatora, M. C., & Ariasi, N. (2013). An eye-tracking study of learning from science text with concrete and abstract illustrations. *The Journal of Experimental Education*, *81*(3), 356-384.

Mattingly, I. G. (1992). Linguistic awareness and orthographic form. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 11-26). Amsterdam, The Netherlands: Elsevier.

Mayer, R. E. (2001). *Multimedia learning*. Cambridge: Cambridge University Press.

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction*, *20*(2), 167-171.

Miller, S. V. (2018). Mixed effects modeling tips: Use a fast optimizer, but perform optimizer checks. http://svmiller.com/blog/2018/06/mixed-effects-models-optimizer-checks/

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart and Winston, Inc.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Miyake, A., Emerson, M. J., & Friedman, N. P. (2000). Assessment of executive functions in clinical settings: Problems and recommendations. *Seminars in Speech and Language*, *21*, 169-183.

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*, 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*, 49-100.

Miyake, A., Friedman, N. P., Rettinger, D. A., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621-640.

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology*, *81*, 111-121.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61–64.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Neelon, B., & O'Malley, A. J. (2019). Two-part models for zero-modified count and semicontinuous data. In A. Levy, S. Goring, C. Gatsonis, B. Sobolev, E. van Ginneken, R. Busse (Eds.), *Health Services Evaluation* (pp. 695-716). Springer.

Neelon, B., O'Malley, A. J., & Smith, V. A. (2016a). Modeling zero-modified count and semicontinuous data in health services research part 1: Background and overview. *Statistics in Medicine*, *35*(27), 5070-5093.

Neelon, B., O'Malley, A. J., & Smith, V. A. (2016b). Modeling zero-modified count and semicontinuous data in health services research part 2: Case studies. *Statistics in Medicine*, *35*(27), 5094-5112.

Noland, R. B., Weiner, M. D., Gao, D., Cook, M. P., & Nelessen, A. (2017). Eye-tracking technology, visual preference surveys, and urban design: Preliminary evidence of an effective methodology. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, *10*(1), 98-110.

Nuthmann, A. (2017). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psychonomic Bulletin & Review*, *24*(2), 370-392.

Olsthoorn, N. M., Andringa, S., & Hulstijn, J. H. (2014). Visual and auditory digit-span performance in native and non-native speakers. *International Journal of Bilingualism*, *18*(6), 663-673.

Ostrosky-Solís, F., & Lozano, A. (2006). Digit span: Effect of education and culture. *International Journal of Psychology*, *41*, 333-341.

Oswald, F. L., Mcabee, S. T., Redick, T. S., & Hambrick, D. Z. (2014). The development of a short domain-general measure of working memory capacity. *Behavior Research*, *47*(4), 1343-1355.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive Load Theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1-4.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive Load Theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science*, *32*, 1-8.

Pashler, H., Bain, P., Bottage, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalf, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, US Department of Education.

Perfetti, C. A., & Liu, Y. (2005). Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing*, *18*(3), 193-210.

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: Some implications of research on Chinese for general theories of reading. *Psychological Review*, *112*(1), 43-59.

Perfetti, C. A., & Zhang, S. (1995). Very early phonological activation in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 24-33.

Polio, C. (2007). A history of input enhancement: Defining an evolving concept. In C. Gascoigne (Ed.), *Assessing the impact of input enhancement in second language education* (pp. 1-18). Stillwater, OK: New Forums Press.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Psychology Software Tools, Inc. (2012). *E-Prime 2.0*. Retrieved from http://www.pstnet.com.

Raghunathan, T. (2015). *Missing data analysis in practice*. Boca Raton: CRC Press.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164-171.

Reeves, C., Schmauder, A. R., & Morris, R. K. (2000). Stress grouping improves performance on an immediate serial list recall task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(6), 1638-1654.

Reinders, H., & Stockwell, G. (2017). Computer-assisted SLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 361-375). New York: Routledge.

Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, (62), 93-132.

Rizopoulos, D. (2020). GLMMadaptive: Generalized linear mixed models using adaptive gaussian quadrature. https://CRAN.R-project.org/package=GLMMadaptive

Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition*, *35*(2), 213-235.

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on incidental second language learning: A replication and extension of Reber, Walkenfield, and Hernstadt (1991). In P. Robinson & P. Skehan (Eds.), *Individual differences and second language instruction* (pp. 211-265). Philadelphia: Benjamins.

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 631-678). New York: Blackwell.

Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, *27*, 235-268.

Robinson, P., Mackey, A., Gass, S., & Schmidt, R. (2012). Attention and awareness in second language acquisition. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 247-267). New York: Routledge.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*, 207-231.

Ruiz, S., Rebuschat, P., & Meurers, D. (2019). The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research*, 1-30.

Russell, V. (2012). Learning complex grammar in the virtual classroom: A comparison of processing instruction, structured input, computerized visual input enhancement, and traditional instruction. *Foreign Language Annals*, *45*(1), 42-71.

Sagarra, N. (2007). From CALL to face-to-face interaction: The effect of computer-delivered recasts and working memory on L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 229-248). Oxford: Oxford University Press.

Scheiter, K., & van Gog, T. (2009). Introduction using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources. *Applied Cognitive Psychology*, *23*, 1209-1214.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129-158.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1-64). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Schmidt, R. (2012). Attention, awareness, and individual differences in language learning. In W. M. Chan, K. N. Chin, S. Bhatt, & I. Walker (Eds.), *Perspectives on individual characteristics and foreign language education* (pp. 27-50). Boston, MA: De Gruyter Mouton.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). Explaining the modality and contiguity effects: New insights from investigating students' viewing behavior. *Applied Cognitive Psychology*, *24*, 226-237.

Schmitt, N. (2008). Instructed second language vocabulary acquisition. *Language Teaching Research*, *12*(3), 329-363.

Service, E., & Craik, F. I. M. (1993). Differences between young and older adults in learning a foreign vocabulary. *Journal of Memory and Language*, *32*, 608-623.

Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign-language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*, 155-172.

Service, E., Simola, M., Metsaenheimo, O., & Maury, S. (2002). Bilingual working memory span is affected by language skill. *European Journal of Cognitive Psychology*, *14*, 383-407.

Shahabi, S. R., Abad, F. J., & Colom, R. (2014). Short-term storage is a stable predictor of fluid intelligence whereas working memory capacity and executive function are not: A comprehensive study with Iranian schoolchildren. *Intelligence*, *44*(1), 134-141.

Sharwood Smith, M. (1981). Consciousness raising and the second language learner. *Applied Linguistics*, *5*, 159-168.

Sharwood Smith, M. (1991). Speaking to many minds: On the relevance of different types of language information for the L2 learner. *Second Language Research*, *7*, 118-132.

Shen, H. H. (2004). Level of cognitive processing: Effects on character learning among non-native learners of Chinese as a foreign language. *Language and Education*, *18*(2), 167-182.

Shen, H. H. (2005). An investigation of Chinese-character learning strategies among non-native speakers of Chinese. *System*, *33*, 49-68.

Shen, H. H. (2010). Imagery and verbal coding approaches in Chinese vocabulary instruction. *Language Teaching Research*, *14*(4), 485-499.

Shen, H. H. (2013). Chinese L2 literacy development: Cognitive characteristics, learning strategies, and pedagogical interventions. *Language and Linguistics Compass*, *7*(7), 371-387.

Shen, H. H., & Ke, C. (2007). Radical awareness and word acquisition among nonnative learners of Chinese. *The Modern Language Journal*, *91*, 97-111.

Silverman, M. J. (2007). The effect of paired pitch, rhythm, and speech on working memory as measured by sequential digit recall. *Journal of Music Therapy*, *44*(4), 415-427.

Sonderegger, M., Wagner, M., & Torreira, F (2018). Quantitative methods for linguistic data. http://people.linguistics.mcgill.ca/~morgan/book/index.html

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, *25*, 293-320.

SR Research Ltd. (2017). *SR Research Experiment Builder 2.1.140*. Retrieve from http://www.sr-research.com/eb.html

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.

Sunderman, G., & Kroll, J. F. (2009). When study-abroad experience fails to deliver: The internal resources threshold effect. *Applied Psycholinguistics*, *30*, 79-99.

Sweller, J. (1999). *Instructional design in technical areas*. Camberwell: ACER Press.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. New York: Springer.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251-296.

Tabachnick, B. G., & Fidell, L. S. (2018). *Using multivariate statistics* (8th ed.). New York: Pearson.

Tamnes, C. K., Walhovd, K. B., Grydeland, H., Holland, D., Østby, Y., Dale, A. M., & Fjell, A. M. (2013). Longitudinal working memory development is related to structural maturation of frontal and parietal cortices. *Journal of Cognitive Neuroscience*, *25*, 1611-1623.

Tavakoli, H. (2012). *A dictionary of research methodology and statistics in applied linguistics*. Tehran: Rahnama Press.

Terraube, J., Helle, P., & Cabeza, M. (2020). Assessing the effectiveness of a national protected area network for carnivore conservation. *Nature Communications*, *11*(1), 1-9.

Truscott, J. (1998). Noticing in second language acquisition: A critical review. *Second Language Research*, *14*, 103-135.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*(6), 1038-1066.

Unsworth, N., Heitz, R. P., Schrock, J., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498-505.

VanPatten, B. (2017). Situating instructed language acquisition: Facts about second language acquisition. *Instructed Second Language Acquisition*, *1*(1), 45-60.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Boca Raton: CRC Press.

van den Noort, M., Bosch, P., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist*, *11*, 289-296.

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2019). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, 1-12.

van Ginkel, J. R., & Kroonenberg, P. M. (2014). Using generalized procrustes analysis for multiple imputation in principal component analysis. *Journal of Classification*, *31*, 242-269.

van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 143-156). New York: Springer.

van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior*, *25*, 325-331.

van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, *20*, 95-99.

van Merrienboer, J. J. G., & Sweller, J. (2005). Cognitive Load Theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147-177.

Wang, M., Perfetti, C. A., & Liu, Y. (2003). Alphabetic readers quickly acquire orthographic structure in learning to read Chinese. *Scientific Studies of Reading*, *7*(2), 127-154.

Wang, S., Zheng, Y., Zheng, C., Su, Y.-H., & Li, P. (2016). An automated test assembly design for a large-scale Chinese proficiency test. *Applied Psychological Measurement*, *40*(3), 233-237.

Warschauer, M. (1996). Computer assisted language learning: An introduction. In S. Fotos (Ed.), *Multimedia language teaching* (pp. 3-20). Tokyo: Logos International.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79-95.

Wechsler, D. (1997). *WAIS-III administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wen, Z. (2015). Working memory in second language acquisition and processing: The Phonological/Executive Model. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 41-62). Bristol: Multilingual Matters.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449-455.

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*, *44*, 573–587.

Williams, J. N., & Lovatt, P. (2003). Phonological memory and rule learning. *Language Learning*, *53*, 67-121.

Williams, J. N. (2012). Working memory and SLA. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 456-471). New York: Routledge.

Williams, J. N. (2015). Working memory in SLA research: Challenges and prospects. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 301–307). Bristol: Multilingual Matters.

Winke, P. M. (2013a). An investigation into second language aptitude for advanced Chinese language learning. *The Modern Language Journal*, *97*(1), 109-130.

Winke, P. M. (2013b). The effects of input enhancement on grammar learning and comprehension. A modified replication of Lee (2007) with eye-movement data. *Studies in Second Language Acquisition*, *35*, 323-352.

Winke, P. M., Gass, S. M., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254-275.

Winke, P. M., Godfroid, A., & Gass, S. M. (2013). Eye-movement recordings in second language research. *Studies in Second Language Acquisition*, *35*, 205-212.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *76*(6), 913-934.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese*. Abingdon, OX: Routledge.

Xu, Y., Chang, L. Y., Zhang, J., & Perfetti, C. A. (2013). Reading, writing, and animation in character learning in Chinese as a foreign language. *Foreign Language Annals*, *46*(3), 423-444.

Xu, Y., Perfetti, C. A., & Chang, L.-Y. (2014). The effect of radical-based grouping in character learning in Chinese as a foreign language. *The Modern Language Journal*, *98*(3), 773-793.

Yeung, A. S. (1999). Cognitive load and learner expertise: Split-attention and redundancy effects in reading comprehension tasks with vocabulary definitions. *The Journal of Experimental Education*, *67*(3), 197-217.

Yeung, A. S., Jin, P., & Sweller, J. (1997). Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. *Contemporary Educational Psychology*, *23*, 1-21.

Yow, W. Q., & Li, X. (2015). Balanced bilingualism and early age of second language acquisition as the underlying mechanisms of a bilingual executive control advantage: Why variations in bilingual experiences matter. *Frontiers in Psychology*, *6*, 1-12.

Zhang, Y., & Li, R. (2016). The role of morphological awareness in the incidental learning of Chinese characters among CSL learners. *Language Awareness*, *25*(3), 179-196.

Zhou, Y. (1986). Modernization of the Chinese language. *International Journal of the Sociology of Language*, *59*, 7-24.

Zuur, A., & Ieno, E. N. (2016). *Beginner's guide to zero-inflated models with R*. Newburgh: Highland Statistics Ltd.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3-14.

Zuur, A. F., Ieno, E. N, Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.