Studies in Second Language Acquisition (2023), 1–23
doi:10.1017/S0272263123000335

RESEARCH ARTICLE

How to present L2 Chinese words effectively for learning: Exploring learning outcomes and learner perceptions

Xuehong (Stella) He[1] and Shawn Loewen[2]

[1]Department of Literature, Media and Language, School of Culture and Communication, Faculty of Humanities and Social Sciences, Swansea University, Singleton Park, Swansea, Wales, UK; [2]Second Language Studies Program, Department of Linguistics, Languages & Cultures, Michigan State University, East Lansing, MI, USA
Corresponding author: Xuehong (Stella) He; Email: xuehong.he@swansea.ac.uk

(Received 11 June 2022; Revised 20 April 2023; Accepted 02 June 2023)

Abstract

Second language (L2) research on input manipulation has focused mainly on increasing the salience of target structures, but presentation formats of L2 input can be another important aspect for manipulation. This study compared the horizontal, vertical, and adjacent formats for presenting the characters, pinyin, and English meaning of L2 Chinese vocabulary, by recruiting 69 English native speakers to study 30 Chinese words in these formats. Learning outcomes were indexed with vocabulary gain scores from pretest to posttest. Learner perceptions of the learning process were recorded with ratings and reasons for preference among these formats. The quantitative results showed the adjacent format generally led to higher gain scores than the other two formats and that L2 proficiency also contributed positively. To learners, the adjacent format was the least preferred, but preference ratings were not associated with gain scores. The qualitative findings suggested format familiarity and layout features as main factors of learner preference.

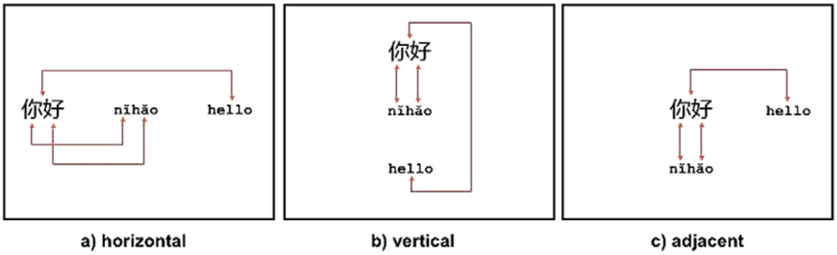
Introduction

Input is essential to second language (L2) acquisition (Gass & Mackey, 2015), and the manipulation of input is a central concern of instructed second language acquisition (ISLA) (Benati, 2016; Lee & Huang, 2008; Loewen, 2020). Based on the assumption that paying attention to input facilitates L2 development (e.g., Schmidt, 2001), L2 research on input manipulation has focused on improving learner attention to targeted linguistic forms (Han et al., 2008; Lee & Huang, 2008; Loewen, 2020) and has proposed pedagogical interventions such as focus on form during meaning-based communication (Long, 1991), including input flood (Hernández, 2011) and input enhancement (Sharwood Smith, 1981). Input manipulation in L2 research has so far mainly investigated enhancing the salience of target structures by increasing the number of exemplars and/or highlighting them in some way (Benati, 2016). Literature in educational psychology suggests that the presentation format of input can be another meaningful aspect for manipulation (Lee & Kalyuga, 2011). Specifically, cognitive load theory (CLT; Sweller et al., 1998, 2019) hypothesizes the split-attention effect for presenting input, affording theoretical foundations with over 30 years of development for exploring effective instructional design to improve human learning (de Bruin & van Merriënboer, 2017; Ginns & Leppink, 2019). Drawing on educational psychology literature, L2 research on input manipulation can further enrich itself and advance its goal of optimizing attention for better learning by investigating presentation formats of input in accordance with general principles of effective instructional design.

Recent CLT review has called for incorporating affective factors to further theoretical development by exploring learners' perceived experience with the instructional design (Plass & Kalyuga, 2019). Similarly, L2 research has highlighted the value of learner perceptions for examining the learning process during a pedagogical intervention (Sato, 2013). This study focuses on both the process and outcome of learning L2 Chinese vocabulary with three different presentation formats (horizontal, vertical, and adjacent; Figure 1), by adopting a mixed-methods approach to combine quantitative results of preference ratings and vocabulary gain scores with qualitative findings of preference reasons. Incorporating educational psychology literature, this study seeks to expand L2 research on input manipulation and support teaching professionals to advance evidence-based L2 vocabulary instruction (He & Godfroid, 2019; He & Loewen, 2022).

Cognitive load theory and the split-attention effect

Cognitive load theory (CLT) is a prominent theory of instructional design in educational psychology (for recent special issues see, e.g., de Bruin & van Merriënboer, 2017; Ginns & Leppink, 2019). It was first proposed to incorporate findings from memory research into developing effective instructional design, and the theory emphasizes a human cognitive architecture that enables learning novel, domain-specific information (Sweller et al., 1998, 2019). Within this framework, novel information is first processed by working memory of limited capacity and duration and then is stored in long-term memory for subsequent retrieval and use (Sweller et al., 1998, 2019). Accordingly, the major goal of instruction is to facilitate transferring novel, domain-specific information from working memory to long-term memory (Sweller et al., 1998, 2019). There are two main categories of cognitive load: intrinsic cognitive load, which depends on the properties of the information as well as the knowledge possessed by the person processing the information, and extraneous cognitive load, which is imposed by



a) horizontal

b) vertical

c) adjacent

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

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

instructional procedures (Sweller et al., 1998, 2019). Because learning is unlikely to happen when the overall cognitive load exceeds the limited working memory capacity, CLT focuses on reducing extraneous cognitive load that originates from nonoptimal instructional procedures, thereby freeing up working memory resources for processing novel information and ultimately, supporting efficient learning (Sweller et al., 1998, 2019).

The split-attention effect is one of the well-established cognitive load effects and provides guidelines for reducing extraneous cognitive load (Sweller et al., 1998, 2019). When multiple sources of information that are essential for understanding but unintelligible in isolation are presented in a separated format, learners need to split their attention to mentally integrate the disparate sources of information, which increases extraneous cognitive load (Ayres & Sweller, 2014). The split-attention effect predicts better learning outcomes with an integrated than a separated format and suggests that learning materials be presented in spatially and/or temporarily integrated formats so as to reduce extraneous cognitive load (Ayres & Sweller, 2014). A meta-analysis of 50 studies with 2,375 novice learners found the split-attention effect is solid and robust regardless of types of effects (spatial vs. temporal), types of presentation (static vs. dynamic), fields of study, or educational levels (Ginns, 2006). In another meta-analysis, Schroeder and Cenkci (2018) focused on the spatial split-attention effect in multimedia learning by including 21 new independent comparisons since Ginns (2006), and their results continued to support the efficacy of integrated over separated formats. Recent research on the split-attention effect has further explored learners' roles and found the benefits of developing learning strategies to physically and mentally integrate materials in separated formats (e.g., de Koning et al., 2020).

Presentation formats in L2 learning

In seeking to provide general principles for instructional design and facilitate learning across fields of study, CLT and the split-attention effect (Sweller et al., 1998, 2019) have important implications for SLA. Similar to SLA theories, CLT assumes the role of working memory in processing novel information and stresses the importance of manipulating input to facilitate attention for efficient learning. The split-attention effect suggests that apart from the salience of target structures (Benati, 2016), presentation formats can afford another meaningful aspect for L2 input manipulation (Lee & Kalyuga, 2011).

Several studies found the effects of presentation formats on L2 learning by changing the location of glosses for reading texts. Yeung et al. (1997) examined 8th graders' reading comprehension and vocabulary learning in L2 English. They created an integrated format by placing the definition near a word in a passage, whereas for the separated format, the word and its definition were placed after the passage. The results showed that compared with the separated format, the integrated format led to better reading comprehension but less vocabulary learning in learners with lower proficiency. Conversely, for more proficient learners, the integrated format resulted in worse reading comprehension but more vocabulary learning than the separated format. Adopting Yeung et al.'s (1997) presentation formats, Yeung (1999) had 5th and 8th graders read L2 English passages corresponding to their grade levels. The results of Yeung (1999) were similar to those in Yeung et al. (1997): 5th graders did better in reading comprehension but worse in vocabulary learning with the integrated format in comparison with the separated format, whereas 8th Graders showed the opposite performance pattern. Yeung et al. (1997) and Yeung (1999) explained that for less

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

proficient learners, the integrated format (in-text definition) saved them from searching and matching the word and its definition, thereby facilitating reading comprehension. On the other hand, the separated format (after-text definition) enabled less proficient learners to focus on the words and thus facilitated vocabulary learning. Differently, for more proficient learners, the in-text definition was redundant for reading comprehension but provided meaningful context for vocabulary learning. Both studies suggested that when presenting learning materials, the split-attention effect should be considered along with task nature and L2 proficiency. In another study, Marefat et al. (2016) compared the effects of in-text and marginal glosses on preintermediate learners' L2 English reading comprehension. For in-text glosses, first language (L1) glosses popped up near a L2 word after it was clicked (integrated format), whereas for marginal glosses, L1 glosses appeared at the right margin (separated format). The integrated format with in-text glosses was found to result in better reading comprehension.

In addition to the location of glosses, the effects of presentation formats on L2 reading have been examined by changing the positions of comprehension questions. Hung (2007) created an integrated format by inserting comprehension questions between paragraphs, whereas in the separated format, the questions were placed after the passage. The results showed better reading comprehension in L2 English for the integrated format. Following Hung's (2007) presentation formats, Al-shehri and Gitsaki (2010) included availability of online dictionaries as an additional factor and investigated reading comprehension and vocabulary learning in L2 English. They found the availability of online dictionaries had a stronger effect on reading comprehension and vocabulary learning than presentation formats. Also using Hung's (2007) presentation formats, Genç and Gülözer (2013) explored the effects of another factor, presentation types (paper-based vs. online), on advanced learners' L2 English reading comprehension. The results showed that comprehension scores were higher in online than paper-based reading but were not significantly different between integrated and separated formats.

Whereas the abovementioned studies focused on L2 English, other research explored presentation formats for learning L2 Chinese. Different from English and other alphabetic writing systems, Chinese generally does not have systematic correspondence between a grapheme (e.g., a letter) and a phoneme (Perfetti et al., 2005). Consequently, pinyin was proposed as a standard phonetic spelling system to facilitate learning Chinese for L1 (Zhou, 1986) and L2 (Everson, 2011) learners. Pinyin adopts English alphabetic letters to spell syllables, but pinyin letters have different pronunciation than English letters (Shen, 2013); in addition, diacritics are used to indicate the five Chinese tones. Generally, learning L2 Chinese vocabulary involves three elements: the shape (characters), the sound (pinyin), and the meaning (Perfetti et al., 2005; Shen, 2013).

Chung (2007) investigated the effects of presentation formats on learning L2 Chinese words by displaying characters, pinyin, and English meaning simultaneously in four different ways: (a) characters-pinyin-English, (b) characters-English-pinyin, (c) English-pinyin-characters, and (d) pinyin-English-characters. The results showed that learning outcomes of the element (pinyin or English) were better when this element was adjacent to characters than when it was distant from characters. Chung explained the finding with the split-attention effect: when the element was far from characters, learners needed to hold the character information in working memory and then search and match it with the element, which may increase extraneous cognitive load and hinder learning.

In another study, Lee and Kalyuga (2011) compared two presentation formats: characters, pinyin, and English meaning were displayed either from left to right in a

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

horizontal format (Figure 1a) or from top to bottom in a vertical format (Figure 1b). Better learning outcomes were found for the vertical rather than the horizontal format. Lee and Kalyuga referred to the split-attention effect as an explanation: in the vertical format with the corresponding pinyin below each character, learners were exempt from holding the character information in working memory for subsequent search and match with the pinyin, thereby reducing extraneous cognitive load. Additionally, Lee and Kalyuga called for future investigation into an adjacent format in which the meaning and pinyin were both next to the characters (Figure 1c), thereby potentially decreasing extraneous cognitive load by reducing the amount of search and match done by learners.

Learner perceptions and instructional practice

In spite of decades of investigation, CLT research has only sparsely examined individual factors other than learners' prior knowledge (Ayres & Paas, 2012). To advance understanding of cognitive processing within the CLT context, researchers have advocated exploring affective factors—namely, how learners perceive or feel during the instructional experience (Ayres & Paas, 2012; Plass & Kalyuga, 2019). L2 researchers have also highlighted the theoretical and pedagogical value of learner perception data, which provide insights into learning processes during pedagogical interventions (Sato, 2013) and inform teachers of learners' responses to L2 teaching practices (Brown, 2009; Hawkey, 2006; Jean & Simard, 2011). It has been suggested that learner perceptions can affect the process and outcome of L2 learning (Grey & Jackson, 2020; Wesely, 2012), and what teachers expect to be effective may or may not be well received by learners (cf. Brown, 2009; Jean & Simard, 2011). Despite the argument that effective L2 pedagogy will work regardless of learner perceptions (Berlin, 2002), learners' negative perceptions of L2 teaching practices might have detrimental effects on their motivation and continuation of L2 learning (Brown, 2009; Jean & Simard, 2011). Consequently, learner perceptions should be explored when considering the efficacy of L2 pedagogy (Sato, 2013) so as to assist teachers to maximize learning outcomes by fostering learner motivation (Brown, 2009; Hawkey, 2006; Jean & Simard, 2011).

Learner perceptions can be investigated both quantitatively and qualitatively (Wesely, 2012). Among the eight L2 studies on presentation formats reviewed above, only four recorded learner perceptions by collecting quantitative difficulty ratings for learning with different formats and/or for completing posttreatment tests (Lee & Kalyuga, 2011; Marefat et al., 2016; Yeung, 1999; Yeung et al., 1997). Notably, these difficulty ratings were mainly used to calculate instructional efficiency scores to compare the efficacy of different formats (see Paas et al., 2003) rather than to provide detailed analysis of learner perceptions. Also, there is a lack of qualitative evidence (e.g., interview data) on how L2 learners perceive different presentation formats.

Research questions

With the goals to advance input manipulation research and evidence-based vocabulary instruction, this study explores both the outcome and the process of learning L2 Chinese words with the horizontal, vertical, and adjacent formats for presenting characters, pinyin, and English meaning (Figure 1). Adopting a mixed-methods approach, we investigated vocabulary gain scores and learner perceptions of the learning process with both quantitative ratings and qualitative reasons for preference. L2 proficiency was

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

included as an additional factor due to its potential effects, as suggested by CLT and previous research findings. Three research questions (RQs) guided this study:

RQ1. How do learners perceive the horizontal, vertical, and adjacent formats for studying L2 Chinese vocabulary?

RQ2. What is the relationship between presentation formats, L2 proficiency, and learning outcomes of L2 Chinese vocabulary?

RQ3. What is the relationship between preference ratings of presentation formats and learning outcomes of L2 Chinese vocabulary?

Method

Participants

The participants were 69 English L1 speakers who did not have Chinese, Korean, or Japanese heritage backgrounds and who had taken elementary Chinese courses in college for less than a year. Generally categorized as novice learners, these participants' Chinese proficiency was further measured by a test described below. According to Nicklin and Vitta's (2021) general guidelines on sample size for instructed L2 vocabulary studies, 57 participants will provide 80% statistical power for three or more counterbalanced repeated-measures analyses with at least a .68 correlation. The current sample of 69 participants with *r* between .826 and .836 for vocabulary gain scores of the three formats was therefore regarded as sufficient.

Materials

Target words

The learning targets were 30 two-character Chinese words chosen from *A Frequency Dictionary of Mandarin Chinese* (Xiao et al., 2009), which covers the 5,004 most commonly used Chinese words based on a 50-million-word corpus of spoken and written texts. All target words were checked to ensure that each character did not appear in the textbooks of the participants' college-level Chinese courses: *Integrated Chinese*, Level 1, Volumes 1 and 2 (Liu et al., 2016, 2017). The words were divided into three 10-word groups, matched in frequency, structural configuration, part of speech, number of strokes, and number of radicals shared with textbook characters (see Online Supplementary Materials A). Three wordlists were then created to counterbalance the presentation formats for all word groups according to a Latin square design. Within each wordlist, all word groups differed in presentation formats, and across wordlists, each word group rotated among the three presentation formats.

Pretest and posttest on vocabulary knowledge

We created a pretest and a posttest with Qualtrics (www.qualtrics.com) to measure knowledge of the form (sound, shape) and meaning of the L2 Chinese vocabulary. Specifically, knowledge of the sound was operationalized as knowledge of the pinyin (Shen & Ke, 2007). Based on Laufer and Goldstein's (2004) bilingual vocabulary test, we developed eight test formats to assess four lexical mappings (see Table 1 for sample items). We also added an "I don't know" option to mitigate the guessing issue that is common in multiple-choice tests (Schmitt et al., 2001).

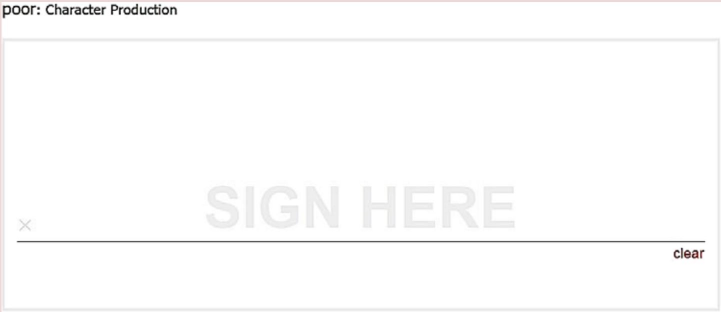
https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Table 1. Sample items of 贫穷 (poor)

Category (a): From meaning to characters and pinyin

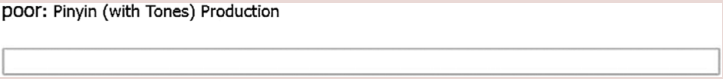
Test Format 1: From meaning to characters – recall

- Write (using the mouse) the Chinese characters for the English meaning


poor: Character Production

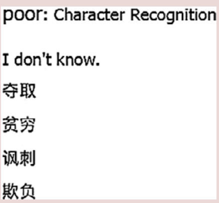
Test Format 2: From meaning to pinyin – recall

- Type the pinyin including tones* of the Chinese characters for the English meaning


poor: Pinyin (with Tones) Production

Test Format 3: From meaning to characters – recognition

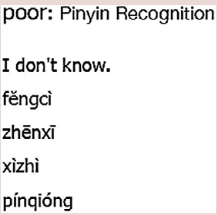
- Choose the Chinese characters for the English meaning


poor: Character Recognition

I don't know.
夺取
贫穷
讽刺
欺负

Test Format 4: From meaning to pinyin – recognition

- Choose the pinyin of the Chinese characters for the English meaning

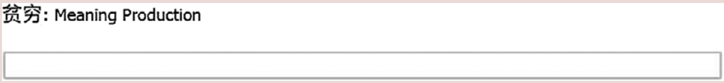

poor: Pinyin Recognition

I don't know.
fěngcì
zhēnxī
xìzhì
pínqióng

Category (b): From characters to meaning and pinyin

Test Format 5: From characters to meaning – recall

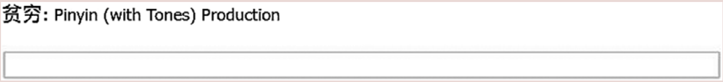
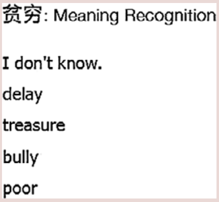
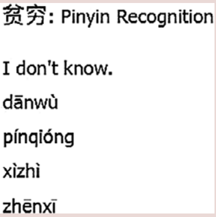
- Type the English meaning for the Chinese characters


贫穷: Meaning Production

(*Continued*)

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Table 1. (*Continued*)

Category (b): From characters to meaning and pinyin
Test Format 6: From characters to pinyin – recall • Type the pinyin including tones* for the Chinese characters  贫穷：Pinyin (with Tones) Production
Test Format 7: From characters to meaning – recognition • Choose the English meaning for the Chinese characters  贫穷：Meaning Recognition I don't know. delay treasure bully poor
Test Format 8: From characters to pinyin – recognition • Choose the pinyin for the Chinese characters  贫穷：Pinyin Recognition I don't know. dānwù pínqióng xìzhì zhēnxī

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

During the pretest, items of the same category (see Table 1) for each word were presented together as a block, with two recall or recognition items—for example, Test Formats 1 and 2—on the same page. After all items of the same category were displayed, those of the other category appeared. That is, after all target words were presented in Test Formats 1 to 4, participants moved to the target words in Test Formats 5 to 8. Participants were not allowed to return to previous pages. The order of test formats corresponded with Nation's (2013) difficulty ranking for productive and receptive knowledge for recall and recognition tests. The pretest and posttest were the same, with the blocks of items randomized within each category for each participant.

Regarding item scoring, all recognition items—Test Formats 3, 4, 7, 8—and the meaning recall items—Test Format 5—received either 0 (incorrect) or 1 (correct) point. Fraction scoring between 0 and 1 was adopted for the form recall items—Test Formats 1, 2, 6. Specifically, for the character recall items (Test Format 1), a correct character received 0.5 points because each word was composed of two characters. For the pinyin recall items (Test Formats 2, 6), $\frac{1}{6}$ points were awarded for a correct tone, pinyin initial, or pinyin final, as each character had one tone, one pinyin initial, and one pinyin final. One example was the pinyin response *jin1qong2* for the word 贫穷 (*poor*). Because the correct answer is *pin2qiong2*, $\frac{1}{6}$ points were given to the correct pinyin final *in* (first character), pinyin initial *q* (second character), and tone *2* (second character), respectively, totaling 0.5 points.

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Table 2. Reliability statistics for test formats

	Test format	Statistics	Index
Recall	(1) From meaning to characters	1.000*	ICC
	(2) From meaning to pinyin	.989*	
	(5) From characters to meaning	.998*	
	(6) From characters to pinyin	1.000*	
Recognition	(3) From meaning to characters	.918	Cronbach's alpha
	(4) From meaning to pinyin	.834	
	(7) From characters to meaning	.938	
	(8) From characters to pinyin	.874	

Note. *$p < .05$.

As the pretest was expected to generate scores approaching zero, test reliability was calculated based on gain scores from pretest to posttest. Following the recommendation of calculating reliability separately for different constructs (Field, 2018), we calculated reliability for each test format (see Table 2), with Cronbach's alpha for the recognition items. For the recall items, two Chinese L1 speakers graded all items separately and intraclass correlation coefficients (ICC) were calculated for interrater reliability. Grading discrepancies were very low (see ICC statistics in Table 2) and were resolved by reaching 100% agreement on the revised grading. All reliability statistics exceeded the recommended benchmark of .70 (Field, 2018).

Chinese proficiency test

A Chinese proficiency test was adapted by selecting four items from each of the four test formats in the reading component of the HSK (Hanyu Shuiping Kaoshi) Level 1 tests. The HSK tests are regarded as the most authoritative standardized Chinese exams for L2 speakers (Wang et al., 2016). For item scoring, one point was awarded for a correct response. Cronbach's alpha was .858 for test reliability.

Postlearning survey and interview

A postlearning, 7-point Likert scale survey was created to collect learners' preferences among the three presentation formats. After participants submitted the survey, their survey responses were displayed on a new webpage for a follow-up audio-recorded interview in which they were asked about their responses, including reasons for their preference ratings. Audio recordings of the interview were then transcribed for analysis. Specifically, the transcription was first generated using the caption function in Kaltura MediaSpace (https://corp.kaltura.com/video-collaboration-communication/enterprise-video-portal/) and then was checked and revised manually by one author.

Procedure

This study adopted a convergent mixed-methods approach to answer the research questions with both quantitative and qualitative analyses (Creswell & Plano Clark, 2018) as well as a within-subject pre/posttest design. Each participant started individually with the Chinese proficiency test and the pretest, then studied the target words and completed the posttest, followed by the postlearning survey and interview. During the learning phase, participants studied the target words in groups of different

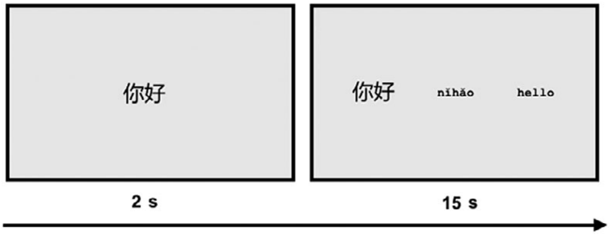


Figure 2. Learning phase.

presentation formats. After studying all words for the first time, they studied these words for a second time. The combinations of word groups and presentation formats were counterbalanced across participants according to a Latin square design. The group order and the word order within each group were randomized for every participant each time.

For the treatment, participants simultaneously saw a word on a computer screen and heard the word being pronounced for 2 seconds (Figure 2). Then, the characters, pinyin, and English meaning appeared simultaneously and stayed on screen for 15 seconds. The total study time for a word was 34 seconds, similar to Lee and Kalyuga's (2011) 30 seconds. The distance between two neighboring elements was 3.5 inches for all presentation formats.

Data analysis

To answer RQ1, quantitative and qualitative analyses were conducted for preference ratings and reasons, respectively. For the preference ratings, as each participant rated three formats (i.e., repeated measures), we calculated descriptive statistics including Cousineau–Morey within-subject confidence intervals (C-M CIs; Baguley, 2012) as well as normality statistics of *z* values for skewness and kurtosis. Because the ratings were normally distributed (see Results), we performed repeated-measures analysis of variance (ANOVA). For preference reasons, thematic analysis (Braun & Clarke, 2012) was conducted. Specifically, the reasons were first grouped into pros and cons for each format, and then main themes were identified and illustrated with representative examples. The number and percentage of participants mentioning each theme were also calculated. Both authors coded 10% of the data separately, with .926 Cohen's kappa indicating good intercoder reliability (McHugh, 2012). Inconsistent coding was discussed and resolved, and one author coded the remaining data.

For RQ2 and RQ3, descriptive statistics including C-M CIs were calculated for the vocabulary gain scores from pretest to posttest for each presentation format (i.e., repeated-measures). Bootstrapped descriptive statistics were calculated for L2 proficiency scores. To initially explore the relationships of learning outcomes to L2 proficiency and format preference, we calculated bootstrapped Pearson's correlations between gain scores and proficiency scores as well as between gain scores and preference ratings.

Recently, L2 researchers (e.g., Cunnings, 2012; Cunnings & Finlayson, 2015; Gries, 2021; Linck & Cunnings, 2015) have recommended mixed-effects modeling for addressing the "language-as-fixed-effect-fallacy" issue (Clark, 1973) by including both participants and language stimuli as random effects in a single analysis

(Baayen et al., 2008). Therefore, to answer RQ2 and RQ3, we conducted mixed-effects modeling for vocabulary gain scores as item-level outcome data. Considering the drawbacks of the model selection approach (Whittingham et al., 2006), we adopted a theoretically and empirically driven approach to choose fixed and random effects. Specifically, for fixed effects, we tested presentation formats and preference ratings as predictors of interest and L2 proficiency scores as a covariate. For random effects, we hypothesized that participants varied in their average gain scores and their extent of being affected by presentation formats and preference ratings and therefore included by-participant random intercept and random slopes for these two predictors. We also assumed that average gain scores and effects of presentation formats varied among words and thus added by-word random intercept and random slope for presentation formats. Notably, convergence problems can occur when the parameters for random effects cannot be reliably estimated by the computational programs, and the random effects component may need to be simplified successively until the model converges (Matuschek et al., 2017). Consequently, we started with the random effects hypothesized above and followed recommended remedies (Brauer & Curtin, 2018; Meteyard & Davies, 2020) to address convergence problems and empirically modify the random effects if necessary.

To select appropriate modeling methods, we checked the distribution of the outcome data (Cunnings & Finlayson, 2015; Zuur et al., 2009). For all recognition and the meaning recall items with 0/1 scoring (Test Formats 3, 4, 5, 7, 8), a binomial distribution was chosen to build a mixed logit model (see Jaeger, 2008). For the form recall items with fraction scoring ([0, 1]; Test Formats 1, 2, 6), all scores were multiplied by 6 to convert to integers (0, 6) in order to avoid infinite numbers. The converted scores contained excessive zeros (96.94%). Hurdle models (also called zero-altered or two-part models) provide an effective tool to model zero-inflated count data by incorporating two submodels: a logit model to account for the occurrence of zeros and nonzeros (binary part) and a truncated Poisson or negative binomial model for the specific values when they are nonzero (positive part; Neelon et al., 2016; Zuur & Ieno, 2016; Zuur et al., 2009). Further exploration of the converted scores showed that overdispersion occurred with zero inflation. The Conway–Maxwell–Poisson distribution provides flexibility in coping with both over- and under-dispersion in the Poisson distribution (Sellers & Premeaux, 2021) and therefore was adopted to build a hurdle mixed-effects model for the form recall items (see Brooks et al., 2017).

We used the glmmTMB function from the glmmTMB package (v1.1.2.3; Brooks et al., 2017) in R (v4.1.2; R Core Team, 2021) for mixed effects modeling. We also used the model_parameters function from the parameters package (v0.17.0; Lüdecke et al., 2020) to calculate 95% CIs for the fixed effects estimates. To explore how each fixed effect influenced overall model fit, we calculated AIC (Akaike information criterion) values for the full model with all fixed effects and for nested models that were identical to the full model except one fixed effect. As smaller AIC values generally suggest better model fit (Matuschek et al., 2017; Meteyard & Davies, 2020), we calculated ΔAIC (AIC change) after excluding each fixed effect by using the full-model AIC value minus the nested-model AIC value. Accordingly, a negative value of ΔAIC indicated better fit of the full model than the nested model that lacked the fixed effect, whereas a positive value suggested better fit of the nested model without the fixed effect. To reduce collinearity among predictors, we conducted grand mean centering (all participants' mean) for L2 proficiency scores as a between-subject variable and group mean centering (each participant's mean) for preference ratings as a within-

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

subject variable (Brauer & Curtin, 2018). Regarding presentation formats as a categorical variable, we conducted deviation coding, which is recommended as generally preferable to treatment coding (Barr et al., 2013) and designated the horizontal format as the reference level. Online Supplementary Materials B provides additional details about the quantitative analyses.

Results

Learner perceptions of presentation formats

Preference ratings

Table 3 reports the descriptive and normality statistics for the preference ratings. The descriptive statistics showed that the average rating was the highest for the horizontal format and the lowest for the adjacent format. Based on the interpretation of non-overlapping C-M 95% CIs for significant differences (Baguley, 2012), the ratings were significantly different among the three formats. The results from the repeated-measures ANOVA also showed significant differences among the ratings of all formats (F = 14.796, $p < .001$). Partial eta squared was 0.179, indicating a large effect size according to Cohen's (1988) benchmarks of effect size (0.01 small, 0.06 medium, 0.14 large). Post hoc tests (see Table 4) further showed that preference ratings were significantly lower for the adjacent format than the horizontal and the vertical formats. Overall, these results suggest that learners preferred the horizontal format the most and the adjacent format the least.

Preference reasons

The analysis of the preference reasons identified format familiarity and layout features as two main factors of learner preference among the three formats. The horizontal format received positive feedback as a familiar format (33 participants, 48%), including being similar to reading from left to right (Excerpt 1) and to the Chinese textbook layout (Excerpt 2). Learners also commended its layout features (7, 10%) such as presenting characters, pinyin, and English meaning in order (Excerpt 3).

Table 3. Descriptive and normality statistics for preference ratings

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

Note. *$p < .01$.

Table 4. Pairwise comparisons between presentation formats

				95% CIs of Mean difference	
I	J	*Mean difference* (I – J)	*p*	*Lower*	*Upper*
Horizontal	Vertical	0.75	.068	–0.40	1.55
Horizontal	Adjacent	2.01*	<.001	1.01	3.02
Vertical	Adjacent	1.26*	.005	0.32	2.20

Note. *$p < .05$, with Bonferroni correction.

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Excerpt 1: *I think horizontal to me is the best because I was taught to read from left to right.*

Excerpt 2: *I like it the most because in our Chinese textbooks that's how it's laid out so it's a lot easier.*

Excerpt 3: *I like the fact that I see the characters and I see that pinyin, and then just the English. I just like how it goes in order like that way, so I just know how it goes.*

The vertical format received both positive and negative comments on format familiarity and layout features. To some learners, this format was still familiar (7, 10%; Excerpt 4), but to others it was less so (10, 14%; Excerpt 5). Regarding layout features, some participants liked the vertical format (8, 12%) because it was in order (Excerpt 6), whereas others disliked it because they felt it was too spaced out (4, 6%; Excerpt 7).

Excerpt 4: *So it is the same order I would normally read and it was just slightly different.*

Excerpt 5: *I usually read from left to right like normal, so when I read up to down, just a little weird.*

Excerpt 6: *So I read the character, pinyin and then the English. So that's very like, you know, in an order.*

Excerpt 7: *I just feel like it's too spaced-out.*

Compared with the other two formats, the adjacent format was less popular due to its unfamiliar format (3, 4%; Excerpt 8). Regarding layout features, it received some positive feedback (15, 22%), such as displaying the three elements close together (Excerpt 9). Nonetheless, many learners reacted negatively toward its layout (32, 46%), with comments such as "weird" (Excerpt 10) and "all over the place" (Excerpt 11).

Excerpt 8: *But like it just kind of threw me off a little bit, you know, because you're just not used to seeing like information presented like that to you.*

Excerpt 9: *I liked it because everything was close, so I didn't have to shift my view off from one part to see it all kind of at the same time.*

Excerpt 10: *Just it was like a weird shape to keep like looking back.*

Excerpt 11: *This is kind of all over the place to me. I just can't really organize it.*

Presentation formats, learner factors, and learning outcomes

Descriptive statistics and Pearson's correlations

Table 5 presents descriptive statistics for vocabulary gain scores. According to the interpretation that nonoverlapping C-M 95% CIs suggest statistically significant differences (Baguley, 2012), the adjacent format resulted in significantly higher gain scores than the other two formats, as indicated by the minimal overlap of the C-M 95% CIs between the adjacent and other formats. Differently, the C-M 95% CIs of the horizontal and vertical formats almost fully overlapped with each other, indicating nonsignificant differences between the gain scores. These results suggest that the adjacent format was associated with better vocabulary learning than the horizontal and vertical formats, which two shared similar learning outcomes. Regarding L2 proficiency, the average score was 12.22 (SD = 3.61) and BCa (bias-corrected and accelerated) 95% CIs of the mean were 11.35 and 13.03.

Table 6 presents results for the bootstrapped Pearson's correlations for vocabulary gain scores. Regarding the correlations with L2 proficiency scores, statistically significant, positive relationships were found for all formats, ranging from .27 for the

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Table 5. Descriptive statistics for vocabulary gain scores

			C-M 95% CIs of Mean	
	Mean	*SD*	*Lower*	*Upper*
Horizontal	20.12	10.46	19.41	20.82
Vertical	20.12	9.85	19.41	20.83
Adjacent	21.49	10.38	20.76	22.21

Table 6. Bootstrapped Pearson's correlations for vocabulary gain scores

			BCa 95% CIs of r	
	r	*p*	*Lower*	*Upper*
L2 Proficiency scores				
Horizontal	.27*	.027	.04	.47
Vertical	.32*	.007	.08	.53
Adjacent	.41*	<.001	.20	.58
Preference ratings				
Horizontal	–.04	.768	–.27	.20
Vertical	.09	.481	–.16	.32
Adjacent	.15	.207	–.07	.37

Note. *p < .05.

horizontal format to .41 for the adjacent format. According to Plonsky and Oswald's (2014) benchmarks for interpreting correlations as effect size (.25 small, .40 medium, .60 large), the effect size of these correlations ranged from small to medium. These results indicate that higher L2 proficiency was linked to better vocabulary learning for all formats. The correlations between gain scores and preference ratings were close to zero and statistically nonsignificant, suggesting that learning outcomes were not directly associated with learner preferences among the three formats.

Mixed-effects models

Table 7 reports the mixed logit model for all recognition and the meaning recall items (Test Formats 3, 4, 5, 7, 8). The results showed that compared with the horizontal format, the adjacent format resulted in significantly better learning outcomes (estimate = 0.21, 95% CIs = [0.02, 0.39], p = .027), and by transforming this estimate in log-odds unit back to odds (Jaeger, 2008), learning with the adjacent format led to 1.23 ($e^{0.21}$) times higher of odds in getting a correct answer. Differently, the vertical format did not result in better learning outcomes than the horizontal format (estimate = 0.05, 95% CIs = [-0.11, 0.21], p = .539). The effect of L2 proficiency was also significant, with better learning outcomes for more proficient learners (estimate = 0.10, 95% CIs = [0.04, 0.16], p < .001), with one unit increase in L2 proficiency scores associated with 1.11 ($e^{0.10}$) times higher odds of getting a correct response. Preference rating was not a significant predictor (estimate = 0.02, 95% CIs = [-0.01, 0.05], p = .183), indicating that learner preference was not associated with learning outcomes. The ΔAIC after exclusion from the full model was -1.3 for presentation formats, 0.2 for preference ratings, and -8.4 for L2 proficiency scores. The negative values of ΔAIC for presentation formats and L2 proficiency scores suggested that including these fixed effects improved overall model fit, whereas the positive value for preference ratings indicated that its inclusion did not

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Table 7. Mixed logit model for recognition and meaning recall items

<table>
<tr><td colspan="2">Formula</td><td colspan="6">Gain score ~ Format + Preference + Proficiency + (1+Format | Participant) + (1+Format | Word)</td></tr>
<tr><td colspan="2"></td><td></td><td></td><td colspan="2"><i>95% CIs of Estimate</i></td><td></td><td></td></tr>
<tr><td colspan="2"><i>Fixed effects</i></td><td><i>Estimate</i></td><td><i>SE</i></td><td><i>Lower</i></td><td><i>Upper</i></td><td><i>z</i></td><td><i>p</i></td></tr>
<tr><td colspan="2">Intercept</td><td>−0.53*</td><td>0.15</td><td>−0.82</td><td>−0.23</td><td>−3.51</td><td>&lt;.001</td></tr>
<tr><td colspan="2">Vertical</td><td>0.05</td><td>0.08</td><td>−0.11</td><td>0.21</td><td>0.61</td><td>.539</td></tr>
<tr><td colspan="2">Adjacent</td><td>0.21*</td><td>0.09</td><td>0.02</td><td>0.39</td><td>2.21</td><td>.027</td></tr>
<tr><td colspan="2">Preference</td><td>0.02</td><td>0.02</td><td>−0.01</td><td>0.05</td><td>1.33</td><td>.183</td></tr>
<tr><td colspan="2">Proficiency</td><td>0.10*</td><td>0.03</td><td>0.04</td><td>0.16</td><td>3.35</td><td>&lt;.001</td></tr>
<tr><td colspan="2"><i>Random effects</i></td><td><i>Variance</i></td><td><i>SD</i></td><td></td><td></td><td></td><td></td></tr>
<tr><td>ID</td><td>Intercept</td><td>0.76</td><td>0.87</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Vertical</td><td>0.13</td><td>0.36</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Adjacent</td><td>0.16</td><td>0.40</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Word</td><td>Intercept</td><td>0.32</td><td>0.57</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Vertical</td><td>0.04</td><td>0.19</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Adjacent</td><td>0.07</td><td>0.26</td><td></td><td></td><td></td><td></td></tr>
</table>

Note. *$p < .05$.

result in better model fit. The results of the ΔAIC and the mixed logit model generally aligned with each other.

Table 8 reports the hurdle mixed-effects model for the form recall items (Test Formats 1, 2, 6). The results of the binary part showed that L2 proficiency score was a significant predictor (estimate = -0.10, 95% CIs = [-0.19, -0.01], *p* = .039), indicating that learners with higher L2 proficiency were more likely to obtain learning gains. Significant effects in the binary part were not found for presentation formats (vertical: estimate = 0.17, 95% CIs = [-0.22, 0.56], *p* = .406; adjacent: estimate = -0.03, 95% CIs = [-0.43, 0.36], *p* = .874) or preference ratings (estimate = -0.04, 95% CIs = [-0.12, 0.05], *p* = .385), suggesting that neither presentation formats nor learner preference was associated with successful learning gains. In the positive part, the results were nonsignificant for presentation formats (vertical: estimate = 0.07, 95% CIs = [-0.06, 0.20], *p* = .304; adjacent: estimate = -0.01, 95% CIs = [-0.14, 0.12], *p* = .857), preference ratings (estimate = 0.00, 95% CIs = [-0.02, 0.03], *p* = .768), or L2 proficiency scores (estimate = 0.02, 95% CIs = [0.00, 0.04], *p* = .090), indicating that when learners successfully obtained learning gains, the amount of gains were not associated with presentation formats, learner preference, or L2 proficiency. The ΔAIC after exclusion from the full model was 5.1 for presentation formats, 3.1 for preference ratings, and -3.1 for L2 proficiency scores. The positive values of ΔAIC change for presentation formats and preference ratings suggested that excluding these fixed effects resulted in better model fit, whereas the negative value of L2 proficiency scores indicated that including it improved overall model fit. The results for ΔAIC were generally in accordance with those of the hurdle mixed-effects model.

Overall, the findings from descriptive statistics, bivariate correlations, and mixed-effects models suggest that better performance in recognizing the meaning and form (characters, pinyin) as well as recalling the meaning was associated with the adjacent format and higher L2 proficiency but not learner preference. As for recalling the form (characters, pinyin), more proficient learners were generally more likely to do better, whereas presentation formats or learner preference did not contribute significantly.

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Table 8. Hurdle mixed effects model for form recall items

Binary part[1]							
Formula	Gain score ~ Format + Preference + Proficiency + (1 | Participant) + (1 | Word)						
				95% CIs of *Estimate*			
Fixed effects		*Estimate*	*SE*	*Lower*	*Upper*	*z*	*p*
Intercept		4.38*	0.25	3.88	4.87	17.23	<.001
Vertical		0.17	0.20	−0.22	0.56	0.83	.406
Adjacent		−0.03	0.20	−0.43	0.36	−0.16	.874
Preference		−0.04	0.04	−0.12	0.05	−0.87	.385
Proficiency		−0.10*	0.05	−0.19	−0.01	−2.07	.039
Random effects		*Variance*	*SD*				
ID	Intercept	1.27	1.13				
Word	Intercept	0.70	0.84				
Positive part							
Formula	Gain score ~ Format + Preference + Proficiency + (1 | Participant) + (1 | Word)						
				95% CIs of *Estimate*			
Fixed effects		*Estimate*	*SE*	*Lower*	*Upper*	*z*	*p*
Intercept		1.31*	0.05	1.21	1.41	26.31	<.001
Vertical		0.07	0.07	−0.06	0.20	1.03	.304
Adjacent		−0.01	0.07	−0.14	0.12	−0.18	.857
Preference		0.00	0.01	−0.02	0.03	0.29	.768
Proficiency		0.02	0.01	0.00	0.04	1.69	.090
Random effects		*Variance*	*SD*				
ID	Intercept	0.02	0.14				
Word	Intercept	0.02	0.14				

Note. *p < .05.

[1] The binary part predicts the probability of zeros (Brook et al., 2017).

Discussion

Drawing on educational psychology literature, this study explored another way of manipulating L2 input by comparing three presentation formats—horizontal, vertical, and adjacent—for learning L2 Chinese vocabulary. We will first discuss the learning outcome to compare the effectiveness of these formats and then will focus on the learning process by examining learner perceptions of these formats and their relationship to the learning outcome. Last, we will provide suggestions on how L2 teaching professionals may incorporate the current findings to create positive learning experiences and enhance L2 Chinese vocabulary development.

Effects of presentation formats on learning L2 Chinese vocabulary

The current results suggest that the adjacent format provided more facilitation to L2 Chinese vocabulary learning than the horizontal and vertical formats in the recognition of meaning and form (characters, pinyin) and in the recall of meaning, generally supporting the split-attention effect predicted by CLT (Sweller et al., 1998, 2019).

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Compared with the other two formats, the adjacent format presented characters, pinyin, and English meaning in a more integrated way (Lee & Kalyuga, 2011): in the horizontal and vertical formats, the pinyin standing in between may have interfered with the characters-and-meaning mapping and increased extraneous cognitive load, whereas in the adjacent format, the adjacency between characters and meaning may have reduced split-attention caused by the pinyin interference, thereby streamlining the mapping between characters and meaning (Chung, 2007). In addition, placing the corresponding pinyin below each character in the adjacent format, in comparison with the horizontal format, may have also reduced split-attention and extraneous cognitive load in that learners would not need to hold the character information for subsequent search and match with the pinyin (Lee & Kalyuga, 2011).

Regarding recalling the form of characters and pinyin, the results indicate that the vertical or the adjacent format did not provide significant advantages over the horizontal format. It is well acknowledged that recalling the form of L2 words is generally more difficult than recalling the meaning or recognizing the meaning or form, and the knowledge for form recall may not develop until later stages of L2 vocabulary learning (Nation, 2013). Given the relatively short learning period (34 seconds per word) in this study, the participants may have not yet developed sufficient vocabulary knowledge to perform form recall successfully. Their overall gain scores for the form recall items were conspicuously low (only about 3% gain scores above zero) and may have led to a floor effect that prevented fully assessing the effectiveness of the three formats on developing the knowledge for form recall. It may be that the advantages of the adjacent and/or vertical format(s) would emerge for performing form recall as vocabulary knowledge develops further. Or it may be that the three formats would be similar in developing the knowledge for form recall.

The results also suggest that higher L2 proficiency generally facilitated learning L2 Chinese vocabulary. This finding is in accordance with CLT's (Sweller et al., 1998, 2019) description about intrinsic cognitive load, which depends on both the properties of the information and the prior knowledge of the learner. Specifically, as learners' prior knowledge increases, the intrinsic cognitive load associated with the novel information will decrease and therefore make available more working memory resources to facilitate efficient learning (Sweller et al., 1998, 2019). The role of L2 proficiency may also explain the different results between this and previous research: whereas the learning outcome was similar for the horizontal and vertical formats in the current study, it was significantly better for the vertical than the horizontal format in Lee and Kalyuga's (2011) study. Different from the current participants who were English native speakers without Chinese, Japanese, or Korean heritage backgrounds, Lee and Kalyuga's participants had family members speaking Chinese at home, which may have resulted in more prior knowledge of L2 Chinese, especially pronunciation. Higher L2 proficiency may have enabled Lee and Kalyuga's participants to better enjoy the advantages of reducing split-attention and extraneous cognitive load brought by the vertical over the horizontal format. For the current participants, placing the corresponding pinyin under each character in the vertical format may have not been sufficient to reduce split-attention and extraneous cognitive load to an extent that could generate detectable learning benefits over the horizontal format.

Learner perceptions of presentation formats and learning L2 Chinese vocabulary

The quantitative results suggest that among the three formats, the adjacent format was the least preferred by L2 learners, whereas the horizontal format was the most

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

preferred. The qualitative findings generally echoed the quantitative results and indicate format familiarity and layout features as two main factors of learner preference. These two factors can be regarded as closely connected, as learners' familiarity with a particular format largely depends on whether they have encountered some of its layout features before. The quantitative and qualitative findings together suggest that L2 learners generally preferred more familiar formats based on their previous experience. Although not specifically about presentation formats, L2 research on pedagogical interventions such as peer interaction (Kuo, 2011) has supported the influence of previous experience on learner perceptions of L2 pedagogy. Specifically, L2 learners are more likely to prefer teaching practices that they have experienced before (Tecedor & Perez, 2021).

The results of preference reasons indicated that the horizontal format was typical in L2 Chinese textbooks and also matched English L1 speakers' common habit of reading from left to right. Reading from top to bottom in the vertical format was considered as parallel to the horizontal format but still less common for English L1 speakers. The adjacent format was regarded as unusual and does not typically appear in English reading materials. Prior experience with the horizontal and probably also the vertical formats may have allowed learners to apply familiar strategies when studying with these formats, as Li (2018) found that L2 learners tended to adopt learning strategies that have been developed from previous learning experience. Differently, the adjacent format may have created confusion and pressured L2 learners to explore new learning strategies for it, as some L2 learners criticized it as being "confusing" and reported lack of available learning strategies. Previous experience may have also influenced individual appreciation of the same layout features, as some L2 learners liked the adjacent format for displaying characters, pinyin, and English meaning closely and/or disliked the vertical format for being spaced out, despite the same distance between two adjacent elements in all formats.

Despite L2 learners' different preferences among the three formats, the results suggest that learner preference did not immediately affect L2 Chinese vocabulary development. Although L2 research has revealed the effects of learner perceptions on L2 development (Wesely, 2012), the effectiveness of L2 pedagogy may not always relate to learner perceptions directly (Berlin, 2002). For instance, Kim and Belcher (2020) found that compared with traditional essay writing, digital multimodal composing was regarded as more enjoyable and effective by L2 learners but it did not result in higher syntactic complexity or accuracy. The current findings that the adjacent format was more effective yet less preferred could provide both encouragement and precaution. It would be encouraging if the adjacent format could have an effect regardless of learner preference. However, it may also encounter learner resistance and affect the learning process negatively, for some L2 learners reported losing interest when studying with it. As negative learner perceptions may cause demotivation and discontinuation of L2 learning in the long term (Brown, 2009; Jean & Simard, 2011), support and training should be offered to L2 learners (Tecedor & Perez, 2021) if the adjacent format is to be adopted.

Incorporating presentation formats to promote L2 Chinese vocabulary learning

In ISLA research, a notable call for strengthening the research–pedagogy link has been forwarded by recent studies on L2 teaching professionals' perceptions of research and practice (Marsden & Kasprowicz, 2017; Sato & Loewen, 2019). Particularly, L2 teachers preferred researchers to focus on issues relevant to teaching and to offer practical tools that they can easily apply in class (Sato & Loewen, 2019), which ISLA researchers are

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

recommended to consider in order to increase the pedagogical influence of their work (Loewen, 2020). Although this study was situated in a laboratory context, we believe that the current findings can afford pedagogical implications that will be relevant to real-life classrooms and thus meaningful to L2 teaching professionals, making an initial step in supporting the development of the research–pedagogy dialogue.

The current results suggest that compared with the horizontal and vertical formats, studying with the adjacent format led to 1.23 times higher of odds in successful performance. Some may argue that such a difference from a laboratory study is not substantial enough for nonlaboratory educational contexts. We acknowledge that this is a small difference, but one point to note is that the current participants had only *34 seconds* to study an unknown Chinese word. In normal learning contexts, L2 learners often spend more time on multiple occasions, and the benefits of the adjacent format might be expected to be more substantial and meaningful in such situations.

Given the general effectiveness of the adjacent format, L2 teaching professionals could consider adopting this format to advance L2 Chinese vocabulary development. For materials writers, although space limits in print textbooks may not allow integrating the adjacent format to display glossaries, this format could be incorporated into electronic materials as additional learning support. In class, L2 teachers could adopt the adjacent format to present L2 vocabulary materials via handouts, PowerPoint slides, and Quizlet flashcards. They could also encourage L2 learners to use this format to create L2 vocabulary flashcards for self-study.

Notably, the uncommonness of the adjacent format may require additional support and training for L2 learners. As in the current results, some L2 learners may show confusion and resistance as they first encounter the adjacent format. One way to address this is to provide a brief explanation about the research findings so as to increase L2 learners' confidence and motivation to use this format (Brown, 2009). For example, L2 teachers can display a new word in the adjacent format and explain to learners that research has found this format beneficial for developing the mappings between characters, pinyin, and meaning. As initial guidance for using this format, L2 teachers can advise learners to start from characters first (Chung, 2007) and connect them to pinyin or to meaning in the order they prefer and finally connect pinyin and meaning. To assist learners in becoming familiar with this format, L2 teachers can encourage them to practice studying with this format for a period (e.g., one month) and to develop their own learning strategies for it (de Koning et al., 2020). L2 teachers can also organize class activities for learners to share the learning strategies they develop for using this format. Nonetheless, if some learners still feel uncomfortable after trying the adjacent format for a while, they should be supported to return to their previous study formats so as not to discourage them from L2 vocabulary learning.

Limitations and future directions

This study has several limitations that need further consideration and investigation. The first one concerns the psycholinguistics of Chinese word recognition. As suggested by one reviewer, the current findings can be discussed in light of visual word recognition theories. One example is the lexical constituency model (Perfetti et al., 2005), which proposes that a Chinese word's identity is specified collectively by its orthography, phonology, and semantics, and the identification of written Chinese words entails retrieving the phonological and semantic information from the orthographic form. A recent review suggests that identifying Chinese written words relies dominantly on the

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

orthography-to-semantics route and minimally on the phonologically mediated route (orthography-to-phonology-to-semantics), different from alphabetic writing systems (Li et al., 2022). Due to constraints, such as space, we did not conduct a separate analysis on vocabulary test items that focused on the mapping from characters to pinyin or meaning (Test Formats 5 to 8)—namely, written word identification. Further analysis of these test items might bring in interesting results for visual word recognition theories of Chinese.

Other limitations are related to the research design. Situated in a strictly controlled laboratory context, the learning phase for one word lasted only 34 seconds and the vocabulary posttest was taken right after learning, which do not simulate common experiences in L2 vocabulary learning and testing. Probably due to such a short learning time, the percentage of the average vocabulary gain score to the total test score (240) was very low (horizontal = 8.38%; vertical = 8.38%; adjacent = 8.95%) and about 97% (87) of the form recall items received zero gain scores. Future research can center on real-life classroom contexts and record L2 vocabulary development longitudinally to explore the effects of presentation formats with increased ecological validity. Last, this study focused on English L1 speakers with relatively low L2 Chinese proficiency. Recruiting learners of other L1s with higher L2 Chinese proficiency can provide a more comprehensive picture on the effects of presentation formats.

Conclusion

The current study compared the horizontal, vertical, and adjacent formats for presenting characters, pinyin, and English meaning of L2 Chinese words. The findings suggest that the adjacent format was generally more effective in developing L2 Chinese vocabulary, and L2 proficiency also contributed positively. Additionally, the quantitative results indicate that the adjacent format was the least preferred by L2 learners, but learner preference did not have an immediate effect on L2 vocabulary development. The qualitative findings of preference reasons suggest format familiarity and layout features as two main factors of learner preference. Based on the results, L2 teaching professionals could consider adopting the adjacent format to present L2 Chinese vocabulary together with guidance and support for L2 learners.

Acknowledgments. This research was supported by a *Language Learning* Dissertation Grant and a NFMLTA-NCOLCTL Research Support Grant awarded to the first author. We express our heartfelt gratitude to Dr. Sihui (Echo) Ke, Dr. Wenying Zhou, Ms. Xuefei Hao, and Dr. Ho-Hsin Huang for their strong support in data collection. We really appreciate the help with data analysis from Dr. Yingzhao Chen and the insightful feedback from Dr. Aline Godfroid, Dr. Susan Gass, and Dr. Xiaoshi Li.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263123000335.

Competing interest. The authors declare none.

References

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*, 356–375.

Ayres, P., & Paas, F. (2012). Cognitive load theory: New directions and challenges. *Applied Cognitive Psychology*, *26*, 827–832.

Ayres, P., & Sweller, J. (2014). The split-attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 206–226). Cambridge University Press.

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.

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

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*, 255–278.

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

Berlin, L. N. (2002). What constitutes effective ESL instruction: Common themes from the voices of the students. *Journal of Intensive English Studies*, *14*, 1–21.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*, 389–411.

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association.

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*, 378–400.

Brown, A. V. (2009). Students' and teachers' perceptions of effective foreign language teaching: A comparison of ideals. *The Modern Language Journal*, *93*, 46–60.

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*, 1–20.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

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

de Bruin, A. B. H., & van Merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction*, *51*, 1–9.

de Koning, B. B., Rop, G., & Paas, F. (2020). Learning from split-attention materials: Effects of teaching physical and mental learning strategies. *Contemporary Educational Psychology*, *61*, Article 101873.

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.

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE.

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). Routledge.

Genç, 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*, 171–182.

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

Ginns, P., & Leppink, J. (2019). Special issue on cognitive load theory: Editorial. *Educational Psychology Review*, *31*, 255–259.

Grey, S., & Jackson, C. (2020). The effects of learners' perceptions and affective factors on L2 learning outcomes. *Canadian Modern Language Review*, *76*, 2–30.

Gries, S. T. (2021). (Generalized linear) Mixed-effects modeling: A learner corpus example. *Language Learning*, *71*, 757–798.

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

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Hawkey, R. (2006). Teacher and learner perceptions of language learning activity. *ELT Journal*, *60*, 242–252.

He, X., & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, *53*, 348–371.

He, X., & Loewen, S. (2022). Stimulating learner engagement in app-based L2 vocabulary self-study: Goals and feedback for effective L2 pedagogy. *System*, *105*, 1–13.

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.

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). Cambridge Scholars Publishing.

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

Jean, G., & Simard, D. (2011). Grammar teaching and learning in L2: Necessary, but boring? *Foreign Language Annals*, *44*, 467–494.

Kim, Y., & Belcher, D. (2020). Multimodal composing and traditional essays: Linguistic performance and learner perceptions. *RELC Journal*, *51*, 86–100.

Kuo, I. C. (2011). Student perceptions of student interaction in a British EFL setting. *ELT Journal*, *65*, 281–290.

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

Li, J. (2018). A resource-oriented functional approach to English language learning. *Canadian Modern Language Review*, *74*, 53–78.

Li, X., Huang, L., Yao, P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology*, *1*, 133–144.

Linck, J. A., & Cunnings, I. (2015). Chapter 8: The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207.

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*, 1119–1141.

Liu, Y., Yao, T., Bi, N.-P., Ge, L., & Shi, Y. (2016). *Integrated Chinese Level 1 Volume 1* (4th ed.). Cheng & Tsui.

Liu, Y., Yao, T., Bi, N.-P., Ge, L., & Shi, Y. (2017). Integrated Chinese Leve *1 Volume 2* (4th ed.). Cheng & Tsui.

Loewen, S. (2020). *Introduction to instructed second language acquisition* (2nd ed.). Routledge.

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). John Benjamins.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*, Article 2445.

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.

Marsden, E., & Kasprowicz, R. (2017). Foreign language educators' exposure to research: Reported experiences, exposure via citations, and a proposal for action. *The Modern Language Journal*, *101*, 613–642.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*, 276–282.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, Article 104092.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Neelon, B., O'Malley, A. J., & Smith, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research part 1: Background and overview. *Statistics in Medicine*, *35*, 5070–5093.

Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal*, *105*, 218–236.

https://doi.org/10.1017/S0272263123000335 Published online by Cambridge University Press

Paas, F., Tuovinen, J., Tabbers, H., & van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*, 63–71.

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*, 43–59.

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review*, *31*, 339–359.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

Sato, M. (2013). Beliefs about peer interaction and peer corrective feedback: Efficacy of classroom intervention. *The Modern Language Journal*, *97*, 611–633.

Sato, M., & Loewen, S. (2019). Do teachers care about research? The research–pedagogy dialogue. *ELT Journal*, *73*, 1–10.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*, 55–88.

Schroeder, N. L., & Cenkci, A. T. (2018). Spatial contiguity and spatial split-attention effects in multimedia learning environments: A meta-analysis. *Educational Psychology Review*, *30*, 679–701.

Sellers, K. F., & Premeaux, B. (2021). Conway–Maxwell–Poisson regression models for dispersed count data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *13*, Article e1533.

Sharwood Smith, M. (1981). Consciousness raising and the second language learner. *Applied Linguistics*, *5*, 159–168.

Shen, H. H. (2013). Chinese L2 literacy development: Cognitive characteristics, learning strategies, and pedagogical interventions. *Language and Linguistics Compass*, *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.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*, 261–292.

Tecedor, M., & Perez, A. (2021). Perspectives on flipped L2 classes: Implications for learner training. *Computer Assisted Language Learning*, *34*, 506–527.

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*, 233–237.

Wesely, P. M. (2012). Learner attitudes, perceptions, and beliefs in language learning. *Foreign Language Annals*, *45*, s98–s117.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*, 1182–1189.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese*. Routledge.

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*, 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.

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*. Highland Statistics.

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*. Springer.

Cite this article: He, X. and Loewen, S. (2023). How to present L2 Chinese words effectively for learning: Exploring learning outcomes and learner perceptions. *Studies in Second Language Acquisition,* 1–23. https://doi.org/10.1017/S0272263123000335