Studies in Second Language Acquisition (2024), **46**, 1413–1438
doi:10.1017/S0272263124000275

REPLICATION STUDY

First language effects on incidental vocabulary learning through bimodal input

A multisite, preregistered, and close replication of Malone (2018)

The TwiLex Group

Yingzhao Chen (National University of Singapore), Jianwu Gao (Capital Normal University), Eva-Maria Hirzinger-Unterrainer (University of Innsbruck), Bronson Hui (University of Maryland, College Park), Benjamin Kremmel (University of Innsbruck), Peilin Li (The Hong Kong Polytechnic University), Shuang Ma (Capital Normal University), Ryo Maie (Tohoku University), Eva Puimège (KU Leuven), John Rogers (The Hong Kong Polytechnic University), and Micheline Wilson (University of Maryland, College Park)
Corresponding author: Bronson Hui; Email: bhui@umd.edu

(Received 30 June 2023; Revised 11 March 2024; Accepted 29 March 2024)

Abstract

Despite accumulating evidence, accounts for the efficacy of reading-while-listening (RWL) in facilitating vocabulary learning are largely unexamined, hindering a thorough understanding of the reasons underlying the usefulness of such bimodal input. In this article, we report a close replication of Malone (2018), purposefully manipulating the participants' native language background to shed light on whether the auditory component in RWL promotes spoken-written form mappings. One hundred and eighty-eight English learners from Austria, Belgium, Hong Kong, and Beijing read or read and listened to four stories containing target words for learning. They completed two surprise vocabulary tests and two assessments of working memory capacity. We only replicated a correlation between working memory capacity and the form recognition test reported in the initial study. Thanks to our manipulation, we discovered an important role of L1 background in the effectiveness of RWL for form recognition knowledge. We discuss the implications for RWL research.

There has been consistent evidence that providing aural support (or reading-while-listening (RWL)) facilitates vocabulary learning through reading, often compared with exposure to unimodal input (e.g., reading-only [RO]) (e.g., Brown et al., 2008; Chen, 2021; Malone 2018; Teng, 2016; Vu & Peters, 2022; Webb & Chang, 2015). An important step researchers should take is to empirically test *why* RWL has such benefits for vocabulary learning. Through a more nuanced understanding of the observed

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

advantages, researchers and teaching professionals can appreciate the strengths and limitations of this form of input, addressing, for instance, for whom and under what circumstances RWL is the best option to promote vocabulary learning.

One approach to achieve this is to engage in replication research (McManus, 2022). We carried out a close replication of Malone (2018), who reported, among other findings, that L2 learners performed better on a form recognition and a meaning connection test after only two exposures in the RWL condition than in the RO condition. By intentionally modifying the first language (L1) background of the participants in the present study, we tested the role of providing audio in strengthening the written and spoken form-form connections of the target words to be learned, which has been argued to be a cause of RWL's benefits for vocabulary gains (Malone, 2018; Webb & Chang, 2012).

Literature review

Vocabulary learning through RWL

RWL has generally been shown to be more effective for vocabulary learning, compared with RO conditions (e.g., Chen, 2021; Malone 2018; Vu & Peters, 2022). This advantage has been found for single word items (e.g., Chen, 2021; Malone, 2018) as well as collocations (e.g., Vu & Peters, 2022; Webb & Chang, 2022). Given the benefits of RWL, researchers have proposed different accounts of why RWL may be advantageous over RO for vocabulary learning. One explanation is that bimodal input supports text processing, such as facilitating text segmentation with the help of prosodic information in the audio. This support, in turn, frees up cognitive resources that can be used to attend to unfamiliar lexical items (e.g., Teng, 2016; Webb & Chang, 2022). In line with this account, Brown et al. (2008) reported interview data in which participants commented that RWL helped with segmenting the text into meaningful chunks (i.e., parsing), which in turn aided with their guessing of word meanings. If the assistance of the audio can spare cognitive resources, these can be diverted to other processing operations conducive to vocabulary learning. Thus, it is reasonable to expect that learners with a lower working memory (WM) capacity can benefit a great deal from RWL because more cognitive resources become available. Indeed, several researchers have argued for the role of WM in the context of vocabulary learning through RWL (e.g., Brown et al., 2008; Teng, 2016). Yet, much of the argumentation remains at the conceptual level, indicating a need to empirically test and lay bare the role of WM in RWL vocabulary learning (see more discussion subsequently).

In relation to WM capacity, it has also been proposed that RWL encourages deeper processing due to its higher cognitive demand as a result of processing input from more than one modality (Chen, 2021; Malone, 2018). However, this account is difficult to reconcile with the idea that bimodal input facilitates processing and frees up WM capacity. It seems plausible that the two accounts hold true in different contexts, e.g., depending on learners' proficiency level or L2 reading skill. For slower readers or lower-proficiency learners, bimodal processing might impose a higher cognitive demand than RO, as these learners might have difficulty adjusting their reading speed to the pace of the auditory input (e.g., Conklin et al., 2020). Therefore, controlling for proficiency is important in gaining understanding of the underlying mechanism of RWL.

A third explanation for the benefits of RWL is that it can help connect spoken and written input. Simultaneously listening to and reading a text might aid in establishing and strengthening the links between the spoken and written form of a word, leading to

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

stronger associative memory links in the growing L2 mental lexicon (e.g., Webb & Chang, 2012). Indeed, there is some indirect evidence in support of this account. In an eye-tracking study by Conklin et al. (2020), the authors argued that when learners read slightly ahead in the RWL condition, this provides "a visual cue of the boundaries of upcoming auditory words" (p. 271). It is reasonable to expect that easier and more accurate identification of word boundaries in the speech stream could promote connections between the spoken and written word forms, and stronger mappings in the growing L2 mental lexicon. At the same time, solid empirical evidence for this account is still lacking.

Role of the L1

Even when one can attribute the effects of RWL, at least partly, to the mapping of spoken and written word forms, this benefit may depend on the level of transparent grapheme-phoneme correspondence in the L1, compared with the L2, including differences in writing systems (e.g., alphabetic vs. logographic). Earlier studies have related the differences in L2 phonological perception to the differences in learners' L1 grapheme-phoneme correspondence. For example, learners of English from a Germanic language background rely more on sublexical information when processing the target language than their Chinese counterparts, whose L1 has an opaque grapheme-phoneme correspondence (Ben-Yehudah et al., 2019; Botezatu, 2023; Wang et al., 2003).

As a result, connecting written and spoken word forms during reading could be an easier task for Germanic language learners of English. In contrast, L1 Chinese speakers could encounter relatively more difficulties accessing phonological information when reading in English because they are limited to relying more on a whole-word processing approach. In the context of learning vocabulary through RWL, it could, therefore, be hypothesized that RWL might be more beneficial for Chinese L1 speakers whose L2 orthographic decoding is more holistic. For these learners, the use of the audio may, therefore, play a more important role in the establishment of orthographic-phonological connections. In the case of Germanic language speakers, on the other hand, audio support could be less crucial because these learners already have access to phonological information due to their better sublexical processing during reading. Indeed, a recent study appears to suggest that RWL comprehension does depend on the learner's native language background (Koh, 2023). In this light, manipulating learners' L1 background allows researchers to empirically test the extent to which the benefits of RWL in vocabulary learning can be attributed to the audio strengthening the connections between spoken and written word forms. At present, the moderating effect of L1 background on incidental vocabulary learning from RWL has not been investigated, despite its theoretical implications.

The role of working memory in bimodal processing: Malone (2018)

To date, theoretical accounts of the beneficial effects of RWL have seldom been tested empirically. An exception is a study by Malone (2018), who investigated the role of WM capacity in vocabulary learning through RWL, or aural enhancement, as the author put it. Eighty learners of English as a second language from multiple L1 backgrounds were randomly assigned to four treatment groups, reading four stories with either two or four exposures to the 32 target words in an RO or RWL condition. Additionally, the participants completed three WM tasks, as well as a cloze test to estimate their general

English proficiency. After the exposure phase, the participants completed a form-recognition task and a form-meaning connection test. Generally, the participants performed above chance in both vocabulary measures, indicating that at least some knowledge was gained after exposure to the new vocabulary only twice, regardless of input condition. In terms of modality, the benefits of RWL were only observed in the two-exposure groups for form recognition. For the form-meaning measure, on the other hand, RWL was found to be more beneficial than RO for both exposure groups. Importantly, the results also confirmed that WM had a stronger correlation with learning gains in the RWL than the RO condition at the level of form recognition. However, WM capacity did not correlate with gains at the level of form-meaning mapping.

Motivation for replicating Malone (2018)

Malone's (2018) study lends itself well to replication due to its methodological rigor and transparency. Timing of the input was controlled so that the RO and RWL groups were exposed to the text for the same amount of time. This, Malone argued, optimized conditions for incidental learning in both groups, by limiting the amount of time learners could process the L2 input. Furthermore, by matching the exposure time across conditions, theoretical accounts for the effects of bimodal processing (e.g., facilitated spoken-to-written form mapping) could be tested while controlling for the effect of word processing time.

More importantly, in terms of scope, Malone (2018) was the first to investigate the combined effects of input frequency, modality, and WM capacity on incidental vocabulary learning. By showing that the role of WM capacity varied across reading conditions, Malone's study was a significant step in empirically testing the claim that audio rendition can free up WM capacity, allowing for deeper processing and better vocabulary learning. However, he also conceded that the moderating effects of learners' L1 background could not be examined due to the heterogeneous sample in his study. As a result, the study could not test the hypothesis that L1-L2 grapheme-phoneme correspondence might moderate the effects of RWL and RO, nor address the theoretical position that RWL can serve to strengthen links between spoken and written L2 word forms.

To continue to refine our understanding of the benefits of RWL, the present study investigated whether the differences between RO and RWL observed in Malone's study are moderated by learners' L1 background. More specifically, our close replication examined the extent to which the differences in grapheme-phoneme correspondence based on different writing system types between the L1 and the L2 affect incidental learning through RO and RWL, to gain a more nuanced insight into the underlying mechanisms driving the benefits of RWL in vocabulary acquisition.

The present study

Following Malone (2018), we examined how (i) the audio support (RWL vs. RO), (ii) the number of exposures (two vs. four times), and (iii) working memory capacity influence learning of vocabulary under incidental conditions, in addition to our new variable, (iv) participants' L1 backgrounds (Chinese vs. Germanic).

The original study asked the following five research questions:

RQ1: Controlling for duration of exposure, to what extent does initial vocabulary learning (form recognition) occur with minimal exposures (2×) to new words during incidental vocabulary learning while reading?

RQ2: Controlling for duration of exposure, to what extent can initial form-meaning connections be made with only two exposures to target words during incidental vocabulary learning while reading?

RQ3: Controlling for duration of exposure, are there differential effects of frequency of exposure with four exposures rather than two during incidental vocabulary learning while reading?

RQ4: Controlling for duration of exposure, to what extent does aural enhancement facilitate incidental vocabulary from two and four exposures?

RQ5: To what extent (if any) does aural enhancement of reading passages influence effects of WM during incidental vocabulary while reading?

In addition to replicating these RQs, we included a research question about the moderating L1 background variable:

RQ6: To what extent do L1s of different levels of grapheme-correspondence influence the effects of frequency, aural enhancement, and WM observed in Malone's study?

For RQs 1–5, we hypothesized that our findings would be similar to those obtained in Malone (2018). Specifically, for form learning, we would find RWL to be more effective than RO for the two-exposure groups, but not the four-exposure groups. For form-meaning connections, we expected to also find greater learning gains in the RWL condition, but without frequency effects. In terms of WM, we hypothesized to find correlations of WM scores with learning gains in the RWL condition, but only for form recognition.

For RQ 6, we developed the following hypotheses: The Chinese learners were expected to benefit more from the audio support during RWL, for both two and four exposures on the form-recognition task. We expected the same pattern for meaning learning. Additionally, we expected that the Chinese learners' greater benefit from RWL would be contingent on their WM capacity, whereas such an interaction might not be observed in learners with Germanic L1s.

The current study is a close replication of Malone (2018), per the definition of Porte and McManus (2019). The variable we manipulated is the participants' L1 background, hence the following differences from the initial study: In Malone (2018), participants spoke a range of L1s, whereas we only recruited L1 speakers of Chinese (Mandarin and Cantonese) and of Germanic languages (German and Dutch). Task instructions were written in both the participants' L1s and English, instead of English alone, as in Malone (2018). In addition to the changes related to the manipulated variable, our data collection was completed online, rather than in person, as in Malone (2018), due to local coronavirus disease 2019 restrictions at the time of data collection. Because online data collection did not allow us to address any participants' questions promptly, we added more explanations of the experimental tasks in the instructions as well as video demonstrations of the tasks. For the WM measures, we were also not able to access the Shapebuilder task and the nonword span task used in Malone (2018). Thus, we decided to exclude the Shapebuilder task and created our own nonword span task based on Linck et al. (2013), which was what Malone (2018) used. Table 1 summarizes the changes made in our replication.

Other than the changes mentioned in Table 1, the current study's materials (i.e., the reading texts, the PowerPoint slides used to present the texts, and the texts' audio

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

Table 1. Comparison between the initial and the current replication study.

	Malone (2018)	Current study
L1 background	Miscellaneous	Chinese (Mandarin & Cantonese), Germanic (German & Dutch)
Data collection	In person	Online
WM measures	Operation span	No change
	Nonword span from Linck et al. (2013)	Nonword span created based on Linck et al. (2013)
	Shapebuilder task	Excluded
Task instructions	Written in participants' L2 (English)	Written in both participants' L1s and L2 (English) With video task demonstration More elaborate than Malone (2018)
Academic background (see supplementary materials on OSF for more details).	University students in Intensive English Programs in the US	University students Participants from the Beijing and Austria sites were mostly English majors
L2 learning context	English as a second language	English as a foreign language

recordings), instruments (i.e., vocabulary posttests and the nonword span task), and procedures (i.e., task order) were exactly the same as reported in Malone (2018). Materials, instruments, instructions to participants, and data analysis were preregistered and are available on the OSF website (https://osf.io/vntra/).

Methods

Experimental conditions

Following Malone (2018), the current study had a 2 × 2 between-participants design. Four experimental conditions were determined based on the provision of audio support and the number of times target vocabulary items appeared (i.e., RWL with two exposures, RWL with four exposures, RO with two exposures, and RO with four exposures). In the RWL conditions, participants read and simultaneously listened to auditory recordings of the texts. In the RO conditions, participants read the texts without audio input. In the two-exposure conditions, target words appeared twice. In the four-exposure conditions, target words appeared four times. The texts were the same across all conditions.

Participants

A total of 188 L2 learners of English were recruited from four research sites (Beijing, China: n = 48; Hong Kong, China: n = 46; Belgium: n = 49; and Austria: n = 45). Multisite recruitment allowed us to include participants representing two different types of L1s: Chinese (Mandarin and Cantonese) and Germanic languages (German and Dutch). Thirteen participants were excluded due to technical issues during the experiment or missing data points (four from the Beijing site, one from the Hong Kong site, six from the Belgium site, and two from the Austria site). The final sample size for each site was $n = 44$ for Beijing, $n = 45$ for Hong Kong, $n = 43$ for Belgium, and $n = 43$ for Austria. This sample size of each L1 group matched that in Malone (2018), i.e., a total of around 80 participants.

The mean age of the included participants was 19.54 (standard deviation [*SD*] = 3.10, range = 15–42), slightly lower than that in Malone (2018) (mean = 23.30, *SD* = 6.20, range = 18–50) due to our attempt to match proficiency levels. In this regard, the mean score for the cloze test was 25.51, comparable to that in the initial study (2018) (i.e., 24.93). However, there were some small differences across the data collection sites in terms of proficiency: Beijing = .54 (*SD* = .13), Hong Kong = .46 (*SD* = .17), Belgium = .49 (*SD* = .23), and Austria = .54 (*SD* = .21). In the main analyses, we added the participants' cloze test scores to adjust for any potential effects of proficiency. Similar to Malone (2018), we randomly assigned the participants on each site to one of the four experimental conditions.

Materials

The materials included 32 target words, four stories, and the audio files of the stories. Each story contained eight target words. All materials used in the current study were obtained through either personal communication with the author of the initial study or the IRIS database (https://www.iris-database.org/). Below is a brief description of the materials. For more details, see Malone (2018).

The target words were low-frequency, bi-syllabic English words (e.g., *ibis*). All target words were concrete nouns and occurred in the object position in a sentence. The target words appeared either twice or four times, depending on the experimental condition. None of the target words were essential for answering comprehension questions about the stories. The target words also did not appear at the beginning or end of the screen during the task.

The stories into which the target words were embedded were between 694 and 773 words long. Ninety-six percent of the words used in the stories were within the 0 to 4,000 lemma frequency band in the *Corpus of Contemporary American English* (Davies, 2008). In the RWL conditions, the stories were accompanied by audio recordings. The audio speed was between 120 and 140 words per minute. As attention checks, eight comprehension questions were inserted for each story. The comprehension questions were given at four points in the stories, with two questions at each point. Adding audio support did not appear to interfere with comprehension, as both RWL and RO groups performed somewhat similarly on the comprehension questions (RO2 = .89, *SD* = .08; RO4 = .88, *SD* = .08; RWL2 = .85, *SD* = .09; RWL4 = .82, *SD* = .09).

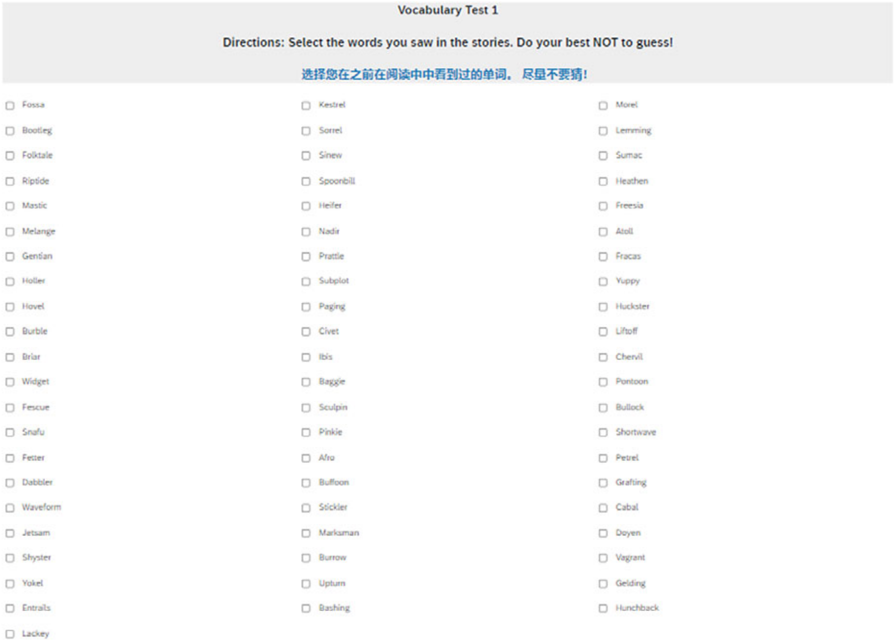
The stories were presented in timed PowerPoint slides in which each line of text would disappear after the audio recording of that line finished playing in the RWL conditions, or after the same amount of time in the RO conditions. In other words, the presentation duration of each line was kept constant in all experimental conditions. The comprehension questions were not timed. The presentation order of the four stories was randomized.

Instruments

Vocabulary posttests

All participants' vocabulary gains were assessed with a form-recognition and a form-meaning connection test, as in Malone (2018). Both tapped into receptive word knowledge. To create an incidental learning condition, participants were not forewarned about the vocabulary posttests. In the form-recognition test, participants selected, from a randomized list of 64 words (32 target words and 32 distractors), the ones that they remembered having seen in the stories. In the form-meaning connection

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press



Vocabulary Test 1

Directions: Select the words you saw in the stories. Do your best NOT to guess!

选择您在之前在阅读中中看到过的单词。 尽量不要猜!

□ Fossa
□ Bootleg
□ Folktale
□ Riptide
□ Mastic
□ Melange
□ Gentian
□ Holler
□ Hovel
□ Burble
□ Briar
□ Widget
□ Fescue
□ Snafu
□ Fetter
□ Dabbler
□ Waveform
□ Jetsam
□ Shyster
□ Yokel
□ Entrails
□ Lackey

□ Kestrel
□ Sorrel
□ Sinew
□ Spoonbill
□ Heifer
□ Nadir
□ Prattle
□ Subplot
□ Paging
□ Civet
□ Ibis
□ Baggie
□ Sculpin
□ Pinkie
□ Afro
□ Buffoon
□ Stickler
□ Marksman
□ Burrow
□ Upturn
□ Bashing

□ Morel
□ Lemming
□ Sumac
□ Heathen
□ Freesia
□ Atoll
□ Fracas
□ Yuppy
□ Huckster
□ Liftoff
□ Chervil
□ Pontoon
□ Bullock
□ Shortwave
□ Petrel
□ Grafting
□ Cabal
□ Doyen
□ Vagrant
□ Gelding
□ Hunchback

Figure 1. Screenshot of the form-recognition posttest for the Beijing site.

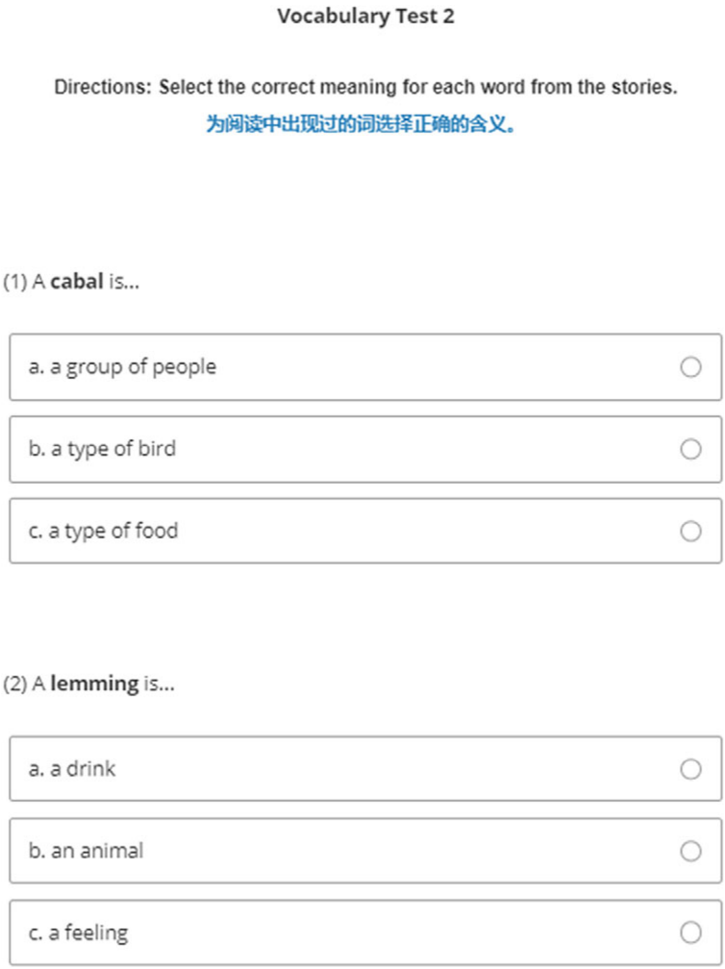
test, participants chose the semantic category out of three choices for each target word. To mimic the paper-and-pencil format of the vocabulary posttests in Malone (2018), all items in both posttests were presented in one single screen. Figures 1 and 2 present screenshots of the posttests. In particular, the online version of the form-recognition test followed Malone (2018) in terms of how the items were ordered and placed. For the form-meaning connection test, participants in the current study could scroll up and down to review all the items in the test.

Following the initial study, each correct answer was worth one point. For the form-recognition test, participants were given one point for correctly identifying each target word, as well as correctly ignoring a distractor. The internal consistency of the test scores based on Kuder-Richardson Formula 20 was. 84 and. 80 for the form recognition and the form-meaning connection test, respectively.

Proficiency measure

Both the current study and Malone (2018) used a cloze test developed by Brown (1980) as a general proficiency measure. The cloze test has been used in numerous studies (e.g., Nekrasova, 2009; Sasayama, 2016) and has been demonstrated to be a reliable measure of English proficiency (e.g., Brown & Grüter, 2022). Participants were asked to fill in 50 blanks in a passage. The passage was displayed in a single display (see Figure 3). In the initial study, Malone adopted what Brown (1980) called the acceptable answer scoring method (Malone, personal communication, June 24, 2022). We obtained the list of acceptable answers from Malone, who had received it from Brown. Participants were given one point if their answer was on the list. Following Malone (2018), minor spelling errors (e.g., *appetite* spelled as *apetite*) were ignored.

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press



Vocabulary Test 2

Directions: Select the correct meaning for each word from the stories.
为阅读中出现过的词选择正确的含义。

(1) A **cabal** is...

a. a group of people ○

b. a type of bird ○

c. a type of food ○

(2) A **lemming** is...

a. a drink ○

b. an animal ○

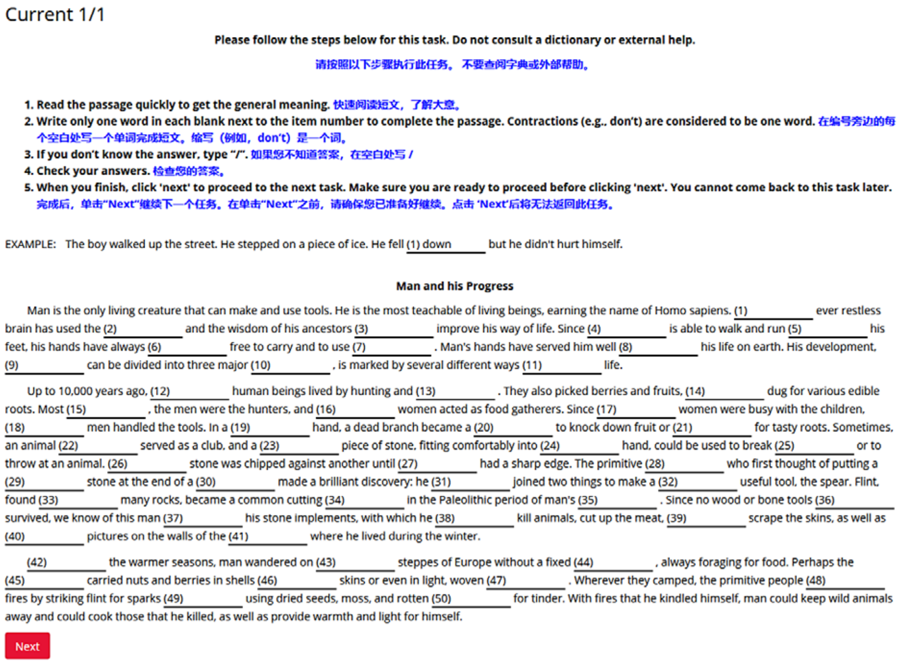
c. a feeling ○

Figure 2. Screenshot of the form-meaning posttest for the Beijing site.

WM measures

Three WM measures were used in Malone (2018): an operation span task, a nonword span task, and the Shapebuilder task. The operation span task in the current study used the same materials and procedures as Malone (2018). We were unable to access the materials used in the nonword span task and in the Shapebuilder task. For the nonword span task, we recreated the materials following the description by Linck et al. (2013), from whom Malone (2018) borrowed the task. We decided to leave out the Shapebuilder task because of unresolvable technical issues in transferring the task to Gorilla, the online data collection platform we used.

In the operation span task, for each trial, participants saw a math equation on the screen, with an uppercase English letter after the equation. Participants were asked to read the math equation aloud and decide whether the equation was correct by pressing one of the two buttons on the screen. After a set of two to five trials, participants were



Please follow the steps below for this task. Do not consult a dictionary or external help.

请按照以下步骤执行此任务。 不要查阅字典或外部帮助。

1. **Read the passage quickly to get the general meaning. 快速阅读短文，了解大意。**
2. **Write only one word in each blank next to the item number to complete the passage. Contractions (e.g., don't) are considered to be one word. 在编号旁边的每个空白处写一个单词完成短文。缩写（例如，don't）是一个词。**
3. **If you don't know the answer, type "/". 如果您不知道答案，在空白处写 /**
4. **Check your answers. 检查您的答案。**
5. **When you finish, click 'next' to proceed to the next task. Make sure you are ready to proceed before clicking 'next'. You cannot come back to this task later. 完成后，单击"Next"继续下一个任务。在单击"Next"之前，请确保您已准备好继续。点击 'Next'后将无法返回此任务。**

EXAMPLE: The boy walked up the street. He stepped on a piece of ice. He fell (1) down ______ but he didn't hurt himself.

Man and his Progress

Man is the only living creature that can make and use tools. He is the most teachable of living beings, earning the name of Homo sapiens. (1) ______ ever restless brain has used the (2) ______ and the wisdom of his ancestors (3) ______ improve his way of life. Since (4) ______ is able to walk and run (5) ______ his feet, his hands have always (6) ______ free to carry and to use (7) ______ . Man's hands have served him well (8) ______ his life on earth. His development, (9) ______ can be divided into three major (10) ______ , is marked by several different ways (11) ______ life.

Up to 10,000 years ago, (12) ______ human beings lived by hunting and (13) ______ . They also picked berries and fruits, (14) ______ dug for various edible roots. Most (15) ______ , the men were the hunters, and (16) ______ women acted as food gatherers. Since (17) ______ women were busy with the children, (18) ______ men handled the tools. In a (19) ______ hand, a dead branch became a (20) ______ to knock down fruit or (21) ______ for tasty roots. Sometimes, an animal (22) ______ served as a club, and a (23) ______ piece of stone, fitting comfortably into (24) ______ hand, could be used to break (25) ______ or to throw at an animal. (26) ______ stone was chipped against another until (27) ______ had a sharp edge. The primitive (28) ______ who first thought of putting a (29) ______ stone at the end of a (30) ______ made a brilliant discovery: he (31) ______ joined two things to make a (32) ______ useful tool, the spear. Flint, found (33) ______ many rocks, became a common cutting (34) ______ in the Paleolithic period of man's (35) ______ . Since no wood or bone tools (36) ______ survived, we know of this man (37) ______ his stone implements, with which he (38) ______ kill animals, cut up the meat, (39) ______ scrape the skins, as well as (40) ______ pictures on the walls of the (41) ______ where he lived during the winter.

(42) ______ the warmer seasons, man wandered on (43) ______ steppes of Europe without a fixed (44) ______ , always foraging for food. Perhaps the (45) ______ carried nuts and berries in shells (46) ______ skins or even in light, woven (47) ______ . Wherever they camped, the primitive people (48) ______ fires by striking flint for sparks (49) ______ using dried seeds, moss, and rotten (50) ______ for tinder. With fires that he kindled himself, man could keep wild animals away and could cook those that he killed, as well as provide warmth and light for himself.

Figure 3. Screenshot of the cloze test for the Beijing site.

prompted to recall the letters they had seen in the set by typing the letters. The scoring of this task was based on the percentage of correctly recalled letters from each set.

For the nonword span task, we first selected 21 phonologically plausible, mono- or bi- syllabic nonwords from the English Lexicon Project (Balota et al., 2007). These 21 nonwords were used repeatedly to form 30 lists, with each list containing seven nonwords. In each trial, participants saw a list of seven nonwords, each nonword appearing on the screen for two seconds. After each trial, participants saw 14 nonwords, half of which they had seen before, and half new to them. Participants indicated whether they had seen the nonword in the trial with a button click.

Procedure

The experiment lasted approximately 2 hours. At the beginning of the experiment, participants saw a welcome page with general information about the experiment. Participants then completed five tasks, in the following order: nonword span task, cloze test, the vocabulary learning task in one of the four conditions, vocabulary posttests, operation span task, and a survey about the participants' demographic information. Data were collected online using Gorilla (https://gorilla.sc/).

Data analysis

Due to space limitations, we do not report all descriptive and inferential statistics, especially when they were not directly relevant to the research questions, such as separate analyses for the two L1 groups. However, those statistics can be found at our OSF page.

We conducted two separate sets of analyses to answer Research Questions 1 through 5. First, we replicated the analytical procedure of Malone (2018), which consisted of two steps. First, we tested Malone's four directional hypotheses regarding how the mean of each reading and exposure condition would be ordered for the form recognition and form-meaning connection tests. We followed the framework of informative hypotheses (Hoijtink, 2011; Hoijtink et al., 2008). This framework takes a confirmatory approach to hypothesis testing in which researchers (only) test statistical hypotheses formulated based on substantive and empirical reasonings. The following are the informative hypotheses tested by Malone:

Form recognition

H_1: $M_{RWL4} = M_{\mathrm{RWL2}} = M_{\mathrm{RO4}} = M_{\mathrm{RO2}}$
H_2: $M_{RWL4} > M_{\mathrm{RWL2}} = M_{\mathrm{RO4}} > M_{\mathrm{RO2}}$
H_3: $M_{RWL4} = M_{\mathrm{RWL2}} = M_{\mathrm{RO4}} > M_{\mathrm{RO2}}$ (supported by Malone)
H_4: $M_{RWL4} = M_{\mathrm{RWL2}} > M_{\mathrm{RO4}} > M_{\mathrm{RO2}}$

Form-meaning connection

H_1: $M_{RWL4} = M_{\mathrm{RWL2}} = M_{\mathrm{RO4}} = M_{\mathrm{RO2}}$
H_2: $M_{RWL4} > M_{\mathrm{RWL2}} > M_{\mathrm{RO4}} > M_{\mathrm{RO2}}$ (supported by Malone)
H_3: $M_{RWL4} > M_{\mathrm{RWL2}} = M_{\mathrm{RO4}} > M_{\mathrm{RO2}}$
H_4: $M_{RWL4} = M_{\mathrm{RWL2}}$ & $M_{\mathrm{RO4}} > M_{\mathrm{RO2}}$

To replicate the analysis, we combined the data from our data collection sites and derived model-based marginal means by fitting a one-way analysis of covariance model to participants' scores from the form recognition (with a scale of 0−64) and form-meaning connection tests (with a scale of 0−32). The model included group (RWL4, RWL2, RO4, RO2) as a between-subjects factor and participants' scores on the cloze test as a covariate. We also included the interaction of group and the cloze test scores to avoid assuming the homogeneity of regression slopes. We compared the four hypotheses based on Bayes factor associated with each hypothesis. In our context, Bayes factor referred to the relative odds of each hypothesis against the other hypotheses combined. We also computed the posterior probability of the hypotheses, which encoded the relative likelihood of the given hypothesis against the others. We used the R package bain (version 0.2.8; Hoijtink et al., 2019) to compute the Bayes factor. The prior distribution supported by the package follows Jeffreys (1961), which is considered to be one of the default non-informative (or objective) prior distributions that work with a wide range of statistical models. We compared our results with those of Malone (2018) by examining which hypothesis is most supported by the data.

In the second step for answering Research Question 5, we computed correlation coefficients to quantify the relationship between participants' vocabulary learning scores and WM scores under the four experimental conditions. Following Malone (2018), we summed scores on the nonword span and the operational span tasks to create single composite WM scores. We used the function correlationBF in the R package BayesFactor (version 0.9.12-4.4: Morey & Rouder, 2011) to compute correlations in a Bayesian framework. Specifically, instead of using confidence intervals and statistical significance, we used 95% credible intervals and the posterior probability of whether the correlation is larger or smaller than 0 to draw statistical inference. Note that there were several influential cases in the participants' WM scores, which could spuriously skew the bivariate relationships between the variables of interest (see our OSF page for illustration). To circumvent this issue, we performed the leave-one-out cross validation on each correlation coefficient. More specifically, we iteratively fit the same correlation by removing one participant from the dataset each time and took the mean of the resulting distribution of correlations as the estimate. We did not preregister this procedure, but it was a necessary step to validate the results. We took our results as replicating Malone (2018) if we observed the same direction and size of the correlations as those reported in the original study.

For RQ6, we investigated the role of L1 backgrounds by fitting a generalized linear mixed model (GLMM) to the form recognition and form-meaning connection data. The dependent variable was participants' accuracy (0 or 1) on each test item, which was regressed on the following predictor variables:

$$\begin{aligned}\text{accuracy} \sim\ & \text{L1} + \text{Frequency} + \text{Condition} + \text{WM} + \text{Cloze} + \text{L1}:\text{Freq} + \text{L1}:\text{Cond} + \\ & \text{Freq}:\text{Cond} + \text{L1}:\text{WM} + \text{Cond}:\text{WM} + \text{L1}:\text{Freq}:\text{Cond} + \text{L1}:\text{Cond}:\text{WM}\end{aligned}$$

- L1: Chinese (0) vs. Germanic (1)
- Frequency: Two (0) vs. four (1) exposures
- Condition: RO (0) vs. RWL (1)
- WM: The composite score of WM (z-score)
- Cloze: The cloze test score (z-score)

Because the dependent variable was on a binary scale, our GLMM was a binomial model that regressed the probability of correct responses on the predictor variables. We used the logit-link function to map the probability (0−1) to the logit of probability (-∞ to +∞) and to model the linear relationship between the dependent and predictor variables. We let the intercept vary among participants and items. Although our experimental design afforded random slopes of the frequency and condition varying among items, we did not incorporate them because it was theoretically implausible to conceive that the effects of the two variables could vary depending on the specific test items.

We estimated the model parameters using Bayesian inference. We used the R-package brms (version 2.18.0; Bürkner, 2017). In Bayesian analysis, prior knowledge in the form of probability distributions is combined with data to generate the posterior distribution of model parameters (Gelman et al., 2013; Norouzian et al., 2018). We took advantage of the statistical results from Malone (2018) to inform our priors (see our preregistration for exact values of the parameters in the prior distributions). We estimated the posterior distributions using the Markov chain Monte Carlo (MCMC) simulation, which consisted of four MCMC chains of 10,000 iterations each, with the first 2,000 iterations not used as a warmup period. We monitored the value of $\widehat{R}$ associated with each parameter to assess whether the MCMC simulation converged on a stable solution (Gelman & Rubin, 1992). We adopted the mean of the posterior distribution as the point estimate, and the highest posterior density interval as the interval estimate of the parameter (i.e. 95% credible intervals). Additionally, we computed the posterior probability of whether a given parameter value is larger or smaller than 0.

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

Results

Testing informative hypotheses (RQs 1–4)

Figure 1 shows the mean and the 95% confidence interval of the participants' scores in each reading and exposure condition. For the form recognition test, the participants in the RWL groups seemed to perform better with four exposures than with two exposures, but this difference was largely driven by the fact that the RWL2 group scored noticeably lower than the three other groups. For the form-meaning connection test, the participants' performances looked overall similar across all conditions. To confirm these observations, we plotted Cohen's *d* for each group comparison as the standardized estimate of the group differences (Figure 2). All the effect sizes were bound within the range of ±0.5, which can be considered small in size, according to the benchmarks suggested by Plonsky and Oswald (2014).

Table 2 summarizes the results of testing the four informative hypotheses on the two vocabulary tests. The table lists four variables: the Bayes factor of a hypothesis against the other hypotheses, the posterior probability of the hypothesis, the winning hypothesis in our dataset, and the winning hypothesis in Malone (2018). On par with our observations (see Figures 1 and 2), we found that H_1 (all groups were equal) was most consistent with our dataset, which indicated that for both vocabulary tests, all experimental groups performed similarly. For the form recognition test, H_1 was 5.14 times more likely than the second hypothesis in rank (H_3) (5.14 = .72/.14), and we found a similar result for the form-meaning connection test: H_1 was 6.46 times more likely than H_4 (6.46 = .84/.13), which was second in rank. Note that we also tested the unconstrained hypothesis (H_u), which did not specify any group differences and hence encoded any other possibilities not represented by our hypotheses. The fact that it was associated with the posterior probability of. 01 for both vocabulary tests reinforces our interpretation that the four experimental groups were more likely to be equal than being different. Hence, our results are not in accordance with those of Malone (2018),

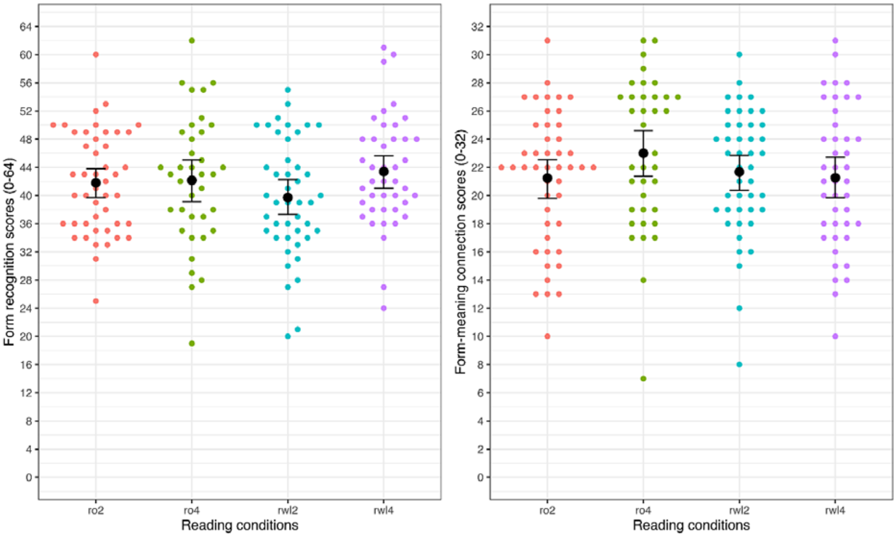


Figure 1. Descriptive summary of participants' performance on the vocabulary tests.
Note: The error bars show 95% confidence intervals.

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

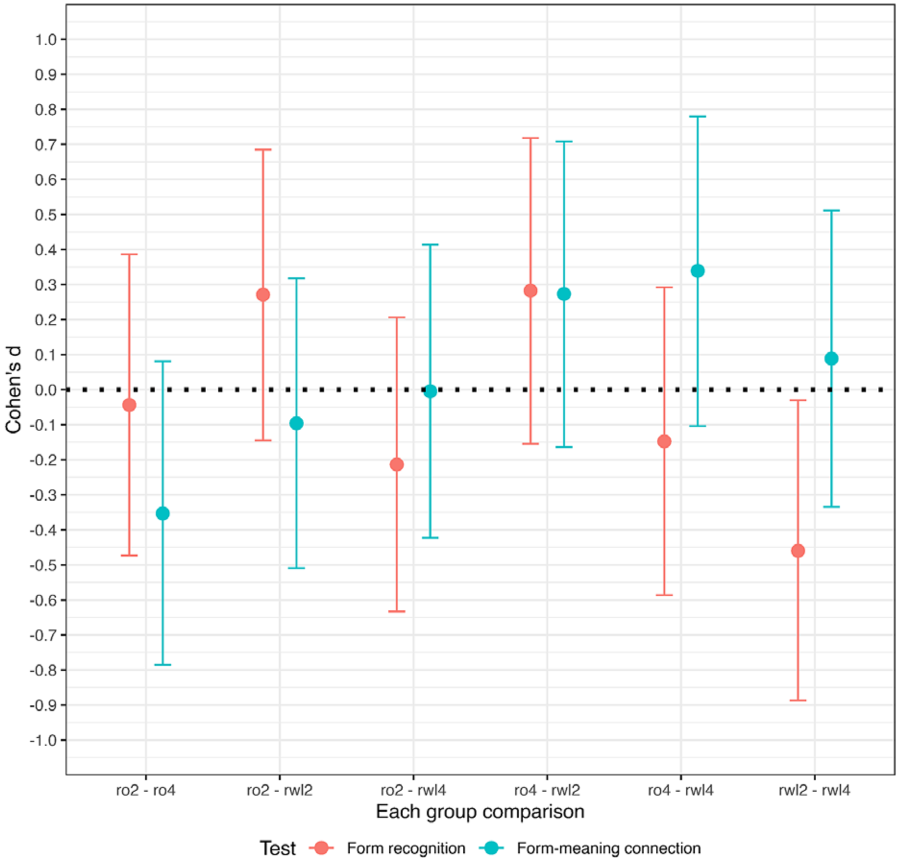


Figure 2. Cohen's *d* of each group difference.
Note: The error bars show 95% confidence intervals.

Table 2. Results of testing informative hypotheses for the vocabulary tests.

	Form recognition				Form-meaning connection			
	BF(H)	*p*(H)	Winner	Malone	BF(H)	*p*(H)	Winner	Malone
H_1	42.41	.72	✓		85.25	.84	✓	
H_2	5.91	.10			0.12	.00		✓
H_3	8.57	.14		✓	1.48	.01		
H_4	0.96	.01			13.47	.13		
H_u		.01				.01		

who found H_3 and H_2 to be most likely for the form recognition and the form-meaning connection tests, respectively. Partly, this finding may be due to the fact that we conflated two different L1 groups (Chinese vs. Germanic L1 speakers), who we hypothesized would react differently to our experimental manipulations. Our Research Question 6 addressed this possibility.

Effects of working memory (RQ5)

Figure 3 shows the correlations between the participants' scores on the two vocabulary tests and their composite scores on the WM tests. For the form recognition test, we also included Malone's (2018) results (i.e., black squares) as they were reported in the initial study. Overall, we replicated Malone in that there were positive correlations between the form recognition scores and WM scores under the RWL condition: r = .30, 95% credible intervals (CrI) [.05,. 50], Pr(r > 0) = .99 for RWL4, and r = .22, 95% CrI [−.04,. 46], Pr(r > 0) = .94 for RWL2. The posterior probability of Pr(r > 0) = .99 and Pr(r > 0) = .94 meant that the correlations were larger than 0 with the probability of. 99 and. 94. Similarly, we also found a positive correlation under the RO condition with two exposures (RO2), r = .29, 95% CrI [.03,. 50], Pr(r > 0) = .98, although the same relationship did not hold with four exposures, r = −.16, 95% CrI [−.42,. 12], Pr(r < 0) = .87. For the form-meaning connection test, there were no notable correlations (r = −.03, 95% CrI [−.33,. 26], Pr(r < 0) = .59 for RWL4, r = .18, 95% CrI [−.12,. 46], Pr(r > 0) = .87 for RO4, and r = −.13, 95% CrI [−.35,. 10], Pr(r < 0) = .85) for RO2) except that we found an unexpected negative correlation under RWL2 (r = −.26, 95% CrI [−.45, −.03], Pr(r > 0) = .98).

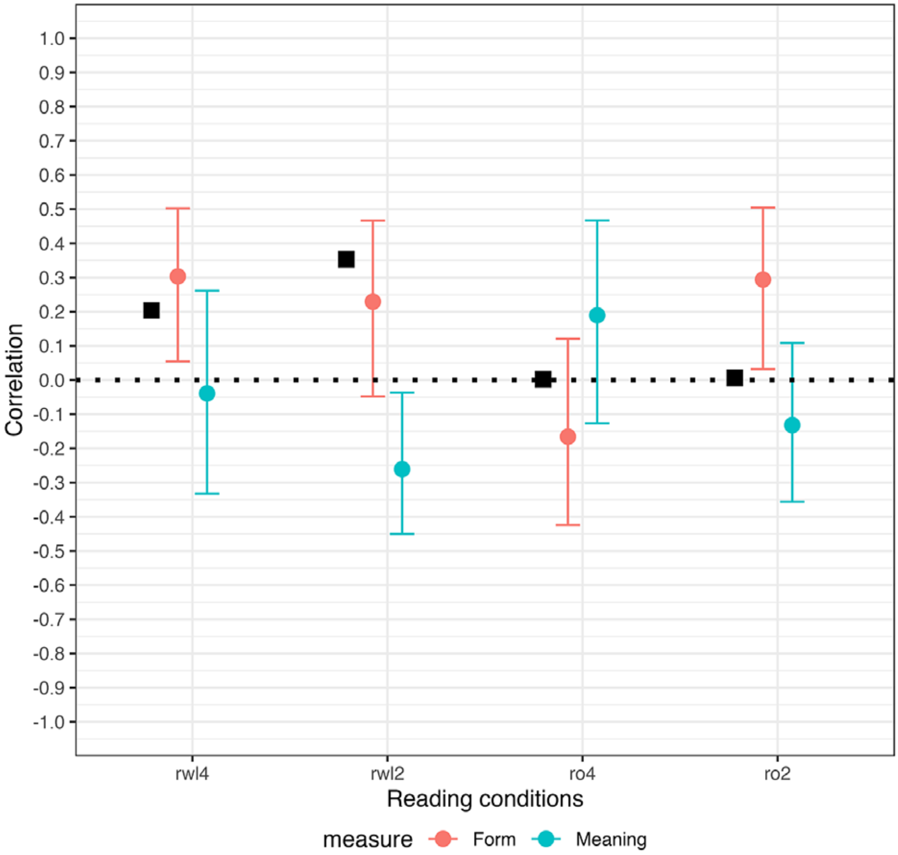


Figure 3. Correlation between the vocabulary tests and working memory scores.
Note: The error bars indicate 95% credible intervals.

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

Effects of first language background (RQ6)

We conducted regression analyses of the participants' scores on the two vocabulary tests to investigate how the effects of reading conditions (RWL vs. RO), the frequency of exposure (four vs. two), and WM capacity changed due to the participants' L1 backgrounds. Table 3 provides the estimate of the model parameters in the binomial GLMM for the form recognition test. The table has three columns: the point estimate of the posterior distribution of the model parameters (Estimate), the interval estimate or 95% CrI of the parameters (95% CrI), and the posterior probability of a given effect being larger or smaller than 0 (Pr[b]), depending on the direction of the point estimate. Note that the posterior distribution of all model parameters converged to a stationary distribution, as the value of $\widehat{R}$ associated with each parameter was within the range of $1.0 \leq \widehat{R} \leq 1.1$ (Gelman & Rubin, 1992).

We found particularly compelling effects of the two-way interaction of reading conditions and WM ($b = 0.32$, 95% CrI [−0.06, 0.69], $\Pr(b > 0) = .95$) and the three-way interaction of L1 backgrounds, reading conditions, and WM ($b = -0.47$, 95% CrI [−0.95, 0.01], $\Pr(b < 0) = .97$). Although less certain, the main effect of L1 backgrounds ($b = -0.32$, 95% CrI [−0.74, 0.11], $\Pr(b < 0) = .92$) and the two-way interaction of L1 backgrounds and frequency ($b = 0.41$, 95% CrI [−0.19, 1.01], $\Pr(b < 0) = .90$) also showed a similar effect in size. To interpret the results, we visually summarized in Figure 4 the model-based marginal mean of each experimental condition (left panel) and the mean of each reading condition as a function of the participants' WM scores (right panel). A close inspection of Figure 4 (right panel) indicates that the three-way interaction of L1 backgrounds, reading conditions, and WM stemmed from the fact that the effect of WM was based on the participants being Chinese L1 speakers and being assigned to the RWL condition. For Chinese L1 participants, the probability of providing correct responses increased from. 54 [.30,. 77] to. 75 [.66,. 82] and to. 80 [.71,. 87] when their WM scores were at the lowest, at the mean, and at the highest point in

Table 3. Parameter estimates from GLMM for the form recognition test.

Fixed effects			
	Estimate	95% CrI	Pr(*b*)
Intercept	1.00	0.51, 1.49	≈ 1.00
L1	−0.32	−0.74, 0.11	.92
Frequency	−0.05	−0.42, 0.40	.58
Condition	0.05	−0.41, 0.52	.58
WM	−0.06	−0.35, 0.24	.65
Cloze	0.34	0.21, 0.47	≈ 1.00
L1: Freq	0.41	−0.19, 1.01	.90
L1: Cond	0.08	−0.50, 0.67	.60
Freq: Cond	0.21	−0.40, 0.80	.74
L1: WM	0.15	−0.21, 0.51	.79
Cond: WM	0.32	−0.06, 0.69	.95
L1: Freq: Cond	−0.38	−1.21, 0.46	.81
L1: Cond: WM	−0.47	−0.95, 0.01	.97

Random effects		
	SD	95% CrI
Random intercepts: Items	1.53	1.29, 1.84
Random intercepts: Participants	0.74	0.64, 0.84

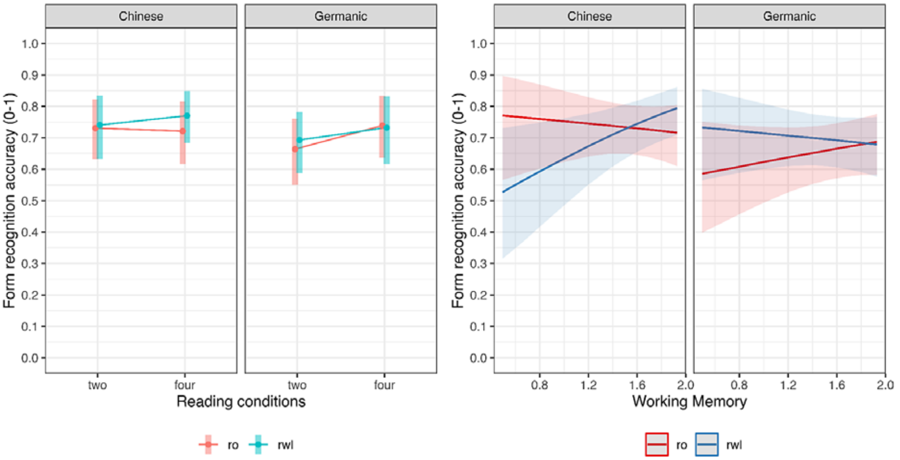


Figure 4. Model-based marginal means of the participants' performance on the form recognition test (*left panel*) and conditional means as a function of working memory scores (*right panel*).

our sample. Because the form recognition test contained 64 items, these probabilities corresponded to answering 34.56 items, 48 items, and 51.2 items correctly. If we assumed the equal accuracy rate across the target and distractor items ($k = 32$ for each category), these probabilities also corresponded to learning the form of 17.28 words, 24 words, and 25.6 words, respectively.

Additionally, one can also interpret the three-way interaction from another perspective; that is, the difference between the RO and RWL conditions was contingent on the participants' WM capacity. Specifically, Chinese-speaking participants in the RO condition scored higher than those in the RWL condition if the participants were of less-than-average WM capacity. This finding suggests that adding audio support may become detrimental to those learners when they do not have an average WM capacity. Lastly, inspecting Figure 4 (left panel), the main effect of L1 backgrounds and the two-way interaction of L1 backgrounds seemed to be driven by the fact that Germanic-speaking participants scored lower than Chinese-speaking participants, especially when they received only two exposures to the target words (regardless of reading conditions).

Table 4 summarizes the estimate of the model parameters in the regression model for the form-meaning connection test. There were only two notable patterns in the results: the main effect of reading conditions ($b = 0.42$, 95% CrI [0.07, 0.77], $\Pr(b > 0) = .99$) and the two-way interaction of the frequency of exposure and reading conditions ($b = -0.50$, 95% CrI [−0.95, −0.05], $\Pr(b < 0) = .98$). Unlike the form recognition test, WM did not seem to predict the participants' scores on the form-meaning connection test.

Figure 5 shows the marginal means of the participants' scores on the form-meaning connection test (left panel) and the conditional means as a function of the participants' WM scores (right panel). Inspecting Figure 5 (left panel), the two-way interaction of the frequency of exposure and reading conditions was driven by the fact regardless of L1 backgrounds, the participants in the RWL condition scored higher than those in the RO condition, but this difference was contingent on the participants receiving only two exposures. Aggregating over L1 backgrounds and controlling WM scores at the mean, the model-based average score of the RWL2 group was. 73 [.67,. 78], whereas that of the

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

Table 4. Parameter estimates from GLMM for the form-meaning connection test.

Fixed effects			
	Estimate	95% CrI	Pr(*b*)
Intercept	0.75	0.39, 1.10	≈ 1.000
L1	−0.04	−0.51, 0.47	.55
Frequency	0.05	−0.27, 0.37	.62
Condition	0.42	0.07, 0.77	.99
WM	0.08	−0.14, 0.29	.76
Cloze	0.03	−0.06, 0.12	.73
L1: Freq	0.17	−0.29, 0.62	.75
L1: Cond	−0.26	−0.71, 0.19	.87
Freq: Cond	−0.50	−0.95, −0.05	.98
L1: WM	−0.08	−0.34, 0.19	.72
Cond: WM	−0.10	−0.38, 0.18	.76
L1: Freq: Cond	0.16	−0.48, 0.79	.69
L1: Cond: WM	−0.04	−0.40, 0.32	.57
Random effects			
		SD	95% CrI
Random intercepts: Items		.60	.44,. 80
Random intercepts: Participants		.33	.22,. 44

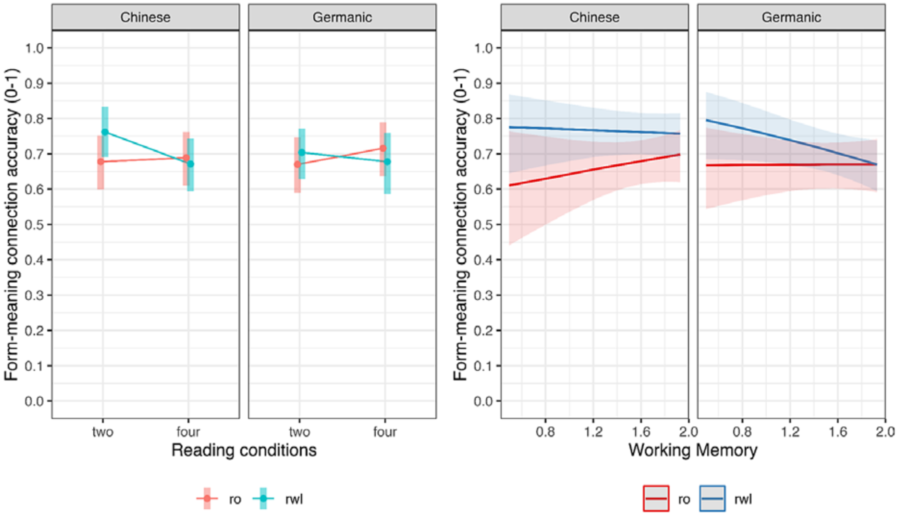


Figure 5. Model-based marginal means of the participants' performance on the form-meaning connection test (*left panel*) and conditional means as a function of WM scores (*right panel*).

RO2 group was. 67 [.61,. 72]. In effect size, this difference (of the probability of. 06) corresponded to learning the meaning of 1.92 words.

Discussion

In this study, we investigated the benefits of RWL on vocabulary learning under incidental conditions by conducting a close replication of Malone (2018). Our research aimed to address two primary objectives. First, we sought to determine the extent to which our results aligned with the initial study regarding the effects of exposure frequency, WM, and aural enhancement on incidental vocabulary learning (addressing RQs 1–5). Second, we aimed to explore the impact of learners' L1 backgrounds, particularly in relation to the grapheme-phoneme correspondence, compared with English (addressing RQ6).

Regarding our first research objective, which aimed to replicate Malone (2018), our results yielded a mixed answer. We hypothesized that we would observe similar outcomes to the initial study, where (i) RWL would be more effective than RO in terms of form recognition after two exposures; (ii) compared with RO, RWL would lead to greater learning gains in form-meaning connections, regardless of whether participants had two or four exposures; and (iii) WM would positively correlate with learning gains in the RWL condition for the form recognition task.

Contrary to our Hypotheses (1) and (2), we found that participants' performance was similar across different reading conditions (RWL vs. RO), regardless of the number of exposures to the new words (two vs. four exposures) for both form-recognition and form-meaning connection tasks. Thus, we were unable to replicate Malone's findings for these two hypotheses (although see discussion of Hypothesis 3 subsequently). However, it is noteworthy that our results may have been influenced by the conflation of two distinct L1 groups (Chinese vs. Germanic L1 learners) in our study. We hypothesized that these two groups would respond differently to our experimental conditions, and the differences could have potentially rendered the effects of reading condition and exposure statistically nonsignificant. In Malone's initial study, although the participants represented a random mix of L1 backgrounds, most of their L1s were Arabic and Chinese, both of which use a writing system distinct from English. It is possible that the participants in the initial study derived greater benefits from the audio enhancement compared with the Germanic learners in our study. We will revisit this point in our discussion of RQ6.

Concerning Hypothesis (3), our results pertaining to the form recognition task largely replicated Malone's findings in three of the four conditions. Specifically, learners with higher WM capacities demonstrated greater learning gains in both the four-exposure RWL (RWL4) and two-exposure RWL (RWL2) conditions, whereas no correlation between WM and learning gains was observed in the four-exposure RO (RO4) condition. Interestingly, we discovered a positive correlation between WM and learning gains in the two-exposure RO (RO2) condition, which was not present in the initial study.

The results that (i) the RWL and RO groups performed similarly and (ii) WM capacities moderated learning in the RWL condition and the RO condition with two target word exposures can be interpreted in at least two ways. First, our findings align with Malone (2018), confirming that RWL places higher demands on learners' WM compared with reading alone. This is in line with previous studies on the online processing of language input. As mentioned earlier, eye-tracking research, such as Conklin et al. (2020), has provided evidence that when new words are presented in the context of familiar words (as operationalized in, e.g., Malone's and the current study), L2 learners may read ahead of the pace of the audio stimuli. In such cases, L2 learners may rely more on the visual text rather than the audio input for overall text comprehension and for understanding of the target stimuli (Webb & Chang, 2012). Consequently, the additional audio input may act as a potential distraction (i.e., "interference"; Kane & Engle, 2000, p. 336), especially for learners with lower WM capacities. On the other hand, learners with higher WM capacities would have the

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

cognitive resources to both avoid the distraction caused by the audio input and effectively use the auditory information to process the words at a deeper level (Craik & Lockhart, 1972; Leow, 2015).

Second, the positive correlation between WM and learning gains in the RO2 condition, but not in the RO4 condition, indicates that the learners may have relied more on their WM during initial exposures. As the frequency of exposure increased, learners with lower WM capacities may have experienced a reduced cognitive load, enabling them to process the new words in ways similar to familiar words, thus diminishing the impact of WM differences (Pellicer-Sánchez, 2016). However, the disparity between our and Malone's findings highlights the complexities in the relationship between cognitive capacities and exposure frequency, which require further investigation.

Although the RWL4 group in our study performed similarly to that in the initial study, our participants in the other three conditions performed considerably better than those in Malone (2018). This could further support the suggestion that even such minimal input leads to traceable beginnings of vocabulary learning, regardless of whether such input is multimodal or not. What remains unclear is why the multimodal input did not lead to more vocabulary gains than the RO condition. There are at least two possible factors that might have played into this. First, the remote data collection and lack of experimental control associated with this study might partly explain why the RWL condition did not generate superior vocabulary gains compared with the RO condition. The nature of the data collection might have meant that participants, particularly those with faster reading speeds, had time to re-read in the RO conditions, whereas participants in the RWL conditions, sticking to the audio speed, might have only had the intended number of exposures (we further discuss learners' cognitive strategies during the training phase subsequently). Second, the nature of the educational background of our samples at the Beijing and Austrian sites, with many participants being English majors, might have contributed to relatively high scores in all conditions, although this potential proficiency advantage was not really reflected in particularly high scores on the proficiency measure.

Our second objective focused on the influence of learners' L1 backgrounds. We hypothesized that RWL might be more beneficial for Chinese L1 speakers, regardless of the frequency of exposure and task, as the audio support could assist these learners in their sublexical processing of the new words, compensating for their reliance on holistic processing that is common with logographic scripts. Conversely, we anticipated that the audio support would be less critical for the Germanic speakers, who already possess phonological information due to their stronger sublexical processing during reading due to their L1 also using a Latin script. Additionally, we expected a stronger correlation between WM and learning outcomes for the Chinese learners compared with Germanic learners.

Regarding the form recognition task, our finding unveils an interplay of L1, WM, and reading condition. This finding is significant, as it not only replicates Malone's findings regarding the positive correlation between WM and aural enhancement but also provides some initial evidence identifying this correlation specifically within the group of Chinese learners. Our findings can be interpreted from at least two perspectives.

First, our results support the notion that the impact of WM is more pronounced or intensified when learners from L1s with an opaque grapheme-phoneme correspondence receive supplementary aural input. Existing research has underscored the influence of L1 backgrounds on word decoding strategies (e.g., Ben-Yehudah et al., 2019). Chinese learners, who are accustomed to a logographic writing system, tend to

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

rely on holistic word processing when decoding L2 English words, even at higher levels of proficiency (e.g., Botezatu, 2023). For this group of learners, the additional aural input could have provided useful sublexical information that they may not have otherwise tapped into. Consequently, their ability to use this information without being distracted (Kane & Engle, 2000) becomes crucial in determining their learning gains. On the other hand, because the Germanic L1 learners already may rely more on sublexical information to process the target language than the Chinese L1 learners, they are less likely to benefit from the additional aural support, making the impact of WM differences less evident.

Viewed from another perspective, our results also indicate that the distraction of additional aural input is only apparent when low WM capacities are coupled with an L1 background characterized by holistic word processing. As illustrated in the right panel of Figure 4, the presence of additional aural input did *not* result in reduced learning gains for Germanic learners with low WM capacities. This finding complements and extends previous studies including Webb and Chang (2012) and Conklin et al. (2020), which attempted to account for the (dis-)advantages of RWL that are found in the literature. Future studies might set out to validate and substantiate this interaction. Further explorations to systematically examine the role of L1 in RWL would also be worthwhile.

On the form-meaning connection task, an interesting finding that diverges from Malone's initial study is a two-way interaction between condition and frequency. Specifically, participants in the RWL condition outperformed those in the RO condition, but this difference was significant only when the new words were encountered twice, with a relatively small average gain of 1.92 more words (at a probability level of. 06). As the frequency of exposures increased, the advantage of RWL diminished. The divergence between our results and Malone's study can be attributed to at least three factors: reactions to the audio recording, differences in sampling contexts, and disparities in cognitive strategies during the training phase.

First, the observed results may be explained by participants' reaction related to the audio recording. As noted earlier, the audio input was played at a fixed speed, which may not have matched the natural reading speed of some participants. Faster readers or individuals with higher proficiency levels (Conklin et al., 2020) might have read ahead in the texts due to the slower pace of the recordings. Moreover, for participants with lower WM capacities, the audio input may have interfered with their ability to process the target words, resulting in poorer performance (see, e.g., Vu & Peters, 2022, Experiment 2). It should be noted that while we used the same recordings as Malone (2018) and recruited participants with similar proficiency levels as those in the initial study, we did not control for individual reading speed or auditory processing (see, e.g., Hui & Godfroid, 2021). This could have contributed to varying levels of reactivity between our study and the initial one. The need for further validation of the audio speed or customization of audio recordings based on individual needs is discussed as a limitation subsequently.

Second, differences in the sampling contexts may also partly explain the disparities in the results. Malone (2018) collected his data in a more controlled laboratory setting, whereas the current study was carried out online. The absence of an experimenter in the unsupervised online setting in our study may have influenced the participants' attention level. Although comprehension tests were incorporated, following the initial study, to screen out inattentive participants (see Procedure section), it is important to recognize that paying attention does not imply engagement in identical cognitive processes.

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

Finally, differences in concurrent cognitive strategies during the training phase may also explain some of the differences in results. There are numerous examples in the SLA literature of participants using explicit strategies (e.g., Robinson, 1997) and developing awareness of the target lexical items (e.g., Godfroid & Schmidke, 2013) under incidental learning conditions. Although the learning conditions were incidental in both Malone's initial study and our experiment, and these incidental conditions may arguably promote incidental learning processes, they ultimately cannot guarantee or control what learning processes occur (see, e.g., Isbell & Rogers, 2021).

Limitations and contributions of this replication

Before proceeding further, it is important to acknowledge several limitations of the current study. First, in our multisite study, we only recruited two specific L1 groups to investigate the role of L1 backgrounds with either logographic or alphabetic writing systems. Second, we replicated the initial study's approach by limiting the number of exposures in our study to two versus four, which may have constrained the conclusions that can be drawn. Additionally, incorporating longitudinal designs in ecologically valid settings in future research would provide a better reflection of how incidental learning naturally occurs and allow for the investigation of additional variables such as input spacing (Webb, 2020). Finally, to gain further insights into the cognitive processes triggered by the instructional conditions, future investigations could incorporate methods such as eye-tracking during the training phase and collect retrospective data through verbal reports, validating the instructional interventions.

Our study, nonetheless, makes several contributions to existing research. First, our study reveals that the impact of WM capacities is intensified in form recognition, when participants come from L1 backgrounds characterized by holistic word-processing strategies. This nuanced finding refines our understanding of aural enhancement by identifying its benefits specifically within a particular type of L1 group, thereby shedding new light on the complexities of the effects of simultaneous input modalities on vocabulary in incidental learning conditions. Pedagogically, this finding suggests that practitioners should take learners' L1 backgrounds into account when incorporating reading while listening into their instruction and provide tailored assistance for learners with lower WM capacities, such as learners with specific learning difficulties (Kormos, 2020), if their L1s have a logographic writing system.

The study further illustrates how replication research can be useful in unpacking the complex nature of the relationship between learners' cognitive capacities and exposure frequencies in learning under incidental learning conditions. In particular, multisite replications, such as this study, may prove beneficial in terms of comparisons of learner groups and contexts, thus enabling evaluations of the generalizability or context-dependence of findings. In this study, recruiting English learners in Europe and Asia made it possible to extend the findings of Malone (2018) by considering the role of different L1 backgrounds on vocabulary learning under different input conditions, as well as study the role of WM in a more nuanced way.

Methodological considerations for future research

To our knowledge, all studies examining the effects of RWL on incidental vocabulary learning have used audio at a fixed speed in their research designs (e.g., Malone, 2018; Teng, 2016; Vu & Peters, 2022; Webb & Chang, 2022; the current study). Although

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

these studies often provide extensive information about the validity and suitability of the reading texts (e.g., proficiency level, tokens, lexical profile), they offer limited details regarding the validity of the audio recordings. Typically, the speech rate (words/minute) is reported, but few studies go beyond this basic information (for an exception, see Webb & Chang, 2022).

From a theoretical standpoint, such design and reporting practices present potential issues that could compromise internal validity. If the audio speeds in previous studies were not optimal for *individual* participants, the results may not accurately reflect the benefits of providing bimodal input for incidental vocabulary acquisition. As we have observed, some participants in our study were negatively affected by the fixed speed of the audio recordings. This finding aligns with the results reported by Vu and Peters (2022, Experiment 2), in which some participants also expressed being distracted by the audio in the RWL condition during retrospective interviews.

Future research investigating RWL might consider a more extensive validation of the audio used in the RWL condition at the *participant* level. This might entail the inclusion of validated audio formats at multiple speeds, where the RWL condition might be tailored to each respective learner. Synchronized textual enhancement might be used in conjunction with bimodal input to encourage participants to follow the audio text more closely (Jung & Lee, 2023). Concurrent measures of attention such as eye-tracking could also be used to track participants' reading to further validate the RWL condition, i.e., the synchrony between the participants' reading and the speed of the audio recording. Finally, the learner's silent reading speed could also be investigated as a moderating variable in that those who read at a similar pace as the audio might benefit differentially from slower and/or faster readers (see Hui, 2024, for an example of RWL comprehension).

Conclusion

The current study replicated Malone (2018) and examined how learners' L1 background may moderate the effects of RWL compared with RO. We failed to replicate Malone's results regarding the comparison of RWL and RO in that our participants in the RWL and RO groups performed similarly in both vocabulary posttests, regardless of the number of times they were exposed to the target words. Our results on the effects of WM aligned with Malone's: There was a positive correlation between WM and form recognition for the RWL group. Most importantly, our findings revealed an interaction among L1 background, WM, and modality for initial form learning: The demand of WM was greater for L1 Chinese learners in RWL conditions. The findings of the current study contribute to the important question of for whom bimodal input may be most beneficial. Based on our findings, researchers may further explore the underlying cognitive mechanism of vocabulary learning through RWL. Pedagogically, this study informs how teachers can use bimodal input, such as through reading aloud learning materials to students, using audiobooks, and text-to-speech functions in e-readers, based on characteristics of materials and students.

Acknowledgments. The authors would like to thank Jonathan Malone, the initial author, for his support of this replication attempt by providing necessary information and materials to make the present study possible.

Credit author statement. **Chen, Yingzhao**: Conceptualization, Methodology, Software, Writing—Original Draft, Writing—Review & Editing, Project Administration; **Gao, Jianwu**: Investigation, Writing—Original Draft, Writing—Review & Editing, Funding Acquisition; **Hirzinger-Unterrainer, Eva**: Investigation; **Hui, Bronson**: Conceptualization, Methodology, Data Curation, Writing—Original Draft, Writing—

https://doi.org/10.1017/S0272263124000275 Published online by Cambridge University Press

Review & Editing, Project Administration, Funding Acquisition; **Kremmel, Benjamin**: Investigation, Writing—Review & Editing, Funding Acquisition; **Li, Peilin**: Investigation; **Ma, Shuang**: Investigation, Writing —Review & Editing; **Maie, Ryo**: Methodology, Formal Analysis, Writing—Original Draft, Writing—Review & Editing, Visualization; **Puimège, Eva**: Investigation, Writing—Review & Editing; **Rogers, John**: Investigation, Writing—Original Draft, Writing—Review & Editing, Funding Acquisition; **Wilson, Micheline**: Writing—Original Draft, Writing—Review & Editing, Project Administration

Data availability statement. The data and materials are available at https://osf.io/vntra/.

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B.,…Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. https://doi.org/10.3758/BF03193014

Ben-Yehudah, G., Hirshorn, E. A., Simcox, T., Perfetti, C. A., & Fiez, J. A. (2019). Chinese-English bilinguals transfer L1 lexical reading procedures and holistic orthographic coding to L2 English. *Journal of Neurolinguistics*, *50*, 136–148. https://doi.org/10.1016/j.jneuroling.2018.01.002

Botezatu, M. R. (2023). The impact of L1 orthographic depth and L2 proficiency on mapping orthography to phonology in L2-English: An ERP investigation. *Applied Psycholinguistics*, *44*, 237–263. https://doi.org/10.1017/S0142716423000176

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal*, *64*, 311–317. https://doi.org/10.2307/324497

Brown, J. D., & Grüter, T. (2022). The same cloze for all occasions? *International Review of Applied Linguistics in Language Teaching*, *60*, 599–624. https://doi.org/10.1515/iral-2019-0026

Brown, R., Waring, B., & Donkaewbua S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163. http://hdl.handle.net/10125/66816

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Chen, Y. (2021). Comparing incidental vocabulary learning from reading-only and reading-while-listening. *System*, *97*, 102442. https://doi.org/10.1016/j.system.2020.102442

Conklin, K., Alotaibi, S., Pellicer-Sánchez, A., & Vilkaitė-Lozdienė (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research*, *36*, 257–276. https://doi.org/10.1177/0267658320921496

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Research*, *11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Davies, M. (2008). The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511. https://doi.org/10.1214/ss/1177011136

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press. https://doi.org/10.1201/b16018

Godfroid, A,. & Schmidke, 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. Yoshika (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 183–205). University of Hawai'i Publishing.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press. https://doi.org/10.1201/b11158

Hoijtink, H., Klugkist, I., & Boelen, P. A. (2008). *Bayesian evaluation of informative hypotheses*. Springer. https://doi.org/10.1007/978-0-387-09612-4

Hoijtink, H., Mulder, J., van Lissa, & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*, 539–556. https://doi.org/10.1037/met0000201

Hui, B. (2024). Scaffolding comprehension with reading-while-listening and the role of reading speed and text complexity. *The Modern Language Journal*. https://doi.org/10.1111/modl.12905

Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*, 1089–1115. https://doi.org/10.1017/S0142716420000193

Isbell, D.R., & Rogers, J. (2021). Measuring implicit and explicit learning and knowledge. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 304–313). Routledge. https://doi.org/10.4324/9781351034784

Jeffrey, H. (1961). *The theory of probability* (3rd ed.). Oxford University Press.

Jung, J. & Lee, M. (2023). Incidental collocational learning from reading-while-listening and the impact of synchronized textual enhancement. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2023-0029

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 336. https://doi.org/10.1037/0278-7393.26.2.336

Koh, J. (2023). Deconstructing the benefits of reading-while-listening on L2 reading comprehension: The influence of cross-orthographic distance. *Foreign Language Annals*. https://doi.org/10.1111/flan.12732

Kormos, J. (2020). Specific learning difficulties in second language learning and teaching. *Language Teaching*, *53*, 129–143. https://doi.org/10.1017/S0261444819000442

Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. Routledge. https://doi.org/10.4324/9781315887074

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R.,…Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*, 530–566. https://doi.org/10.1111/lang.12011

Malone, J. (2018). Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, *40*, 651–675. https://doi.org/10.1017/S0272263117000341

McManus, K. (2022). Are replication studies infrequent because of negative attitudes? Insights from a survey of attitudes and practices in second language research. *Studies in Second Language Acquisition*, *44*, 1410–1423. https://doi.org/10.1017/S0272263121000838

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419. https://doi.org/10.1037/a0024377

Nekrasova, T. M. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language learning*, *59*, 647–686. https://doi.org/10.1111/j.1467-9922.2009.00520.x

Norouzian, R., de Miranda, M., Plonsky, L. (2018), The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68*, 1032–1075. https://doi.org/10.1111/lang.12310

Pellicer-Sánchez, A. (2016). Incidental Ld vocabulary acquisition *from* and *while* reading. *Studies in Second Language Acquisition*, *38*, 97–130. https://doi.org/10.1017/S0272263115000224

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Porte, G., & McManus, K. (2019). *Doing replication research in Applied Linguistics*. Routledge. https://doi.org/10.4324/9781315621395

Robinson, P. (1997). Generalisability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, *19*, 223–247. https://doi.org/10.1017/S0272263197002052

Sasayama, S. (2016). Is a "complex" task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, *100*, 231–254. https://doi.org/10.1111/modl.12313

Teng, F. (2016). Incidental vocabulary acquisition from reading-only and reading-while-listening: A multi-dimensional approach. *Innovation in Language Learning and Teaching*, *12*, 1–15. https://doi.org/10.1080/17501229.2016.1203328

Vu, D. V., & Peters, E. (2022). Learning vocabulary from reading-only, reading-while-listening, and reading with textual input enhancement: Insights from Vietnamese EFL learners. *RELC Journal 53*, 85–100. https://doi.org/10.1177/003368822091148

Wang, M., Koda, K., & Perfetti, A. C. (2003). Alphabetic and nonalphabetic L1 effects in English word identification: A comparison of Korean and Chinese L2 learners. *Cognition*, *87*, 129–149. https://doi.org/10.1016/s0010-0277(02)00232-9

Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–239). Routledge. https://doi.org/10.4324/9780429291586

Webb, S., & Chang, A. (2012). Vocabulary learning through assisted and unassisted repeated reading. *The Canadian Modern Language Review*, *68*, 276–290. https://doi.org/10.3138/cmlr.1204.1

Webb, S., & Chang, A. C.-S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, *37*(4), 651–675. https://doi.org/10.1017/S0272263114000606

Webb, S., & Chang, A. C.-S. (2022). How does mode of input affect the incidental learning of collocations. *Studies in Second Language Acquisition*, *44*, 35–56. https://doi.org/10.1017/S0272263120000297

Cite this article: The TwiLex Group. (2024). First language effects on incidental vocabulary learning through bimodal input: *A multisite, preregistered, and close replication of Malone (2018). Studies in Second Language Acquisition,* 46: 1413–1438. https://doi.org/10.1017/S0272263124000275