University of Reading

Attention and learning in L2 multimodality: a webcam-based eye-tracking study

Article

Published Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Open Access

Zhang, P. ORCID: https://orcid.org/0000-0002-2136-4984 and Zhang, S. (2025) Attention and learning in L2 multimodality: a webcam-based eye-tracking study. Language Learning & Technology, 29 (1). pp. 1-27. ISSN 1094-3501 Available at https://centaur.reading.ac.uk/121795/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.
Published version at: https://hdl.handle.net/10125/73626

Publisher: National Foreign Language Resource Center

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

University of Reading

Reading's research outputs online

Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2025, Volume 29, Issue 1
pp. 1–27

ARTICLE

Attention and learning in L2 multimodality: A webcam-based eye-tracking study

Pengchong Zhang, University of Reading*

Shi Zhang, Chengdu University of Technology

Abstract

Multimodal input can significantly support second language (L2) vocabulary learning and comprehension. However, very little research has examined how L2 learners, especially young learners, allocate attention when exposed to such input and whether learning from multimodal input can be explained by attention allocation. This study therefore investigated individual differences in attention allocation during L2 vocabulary learning with multimodal input and how vocabulary learning and comprehension were influenced by these differences. Forty young learners of French watched two types of multimodal input (Written+Audio+Picture vs. Written+Speaker+Video) and had their eye-movements recorded through online webcam-based eye-tracking technology. They also completed tests of comprehension, vocabulary, and phonological short-term memory (PSTM). We show that greater attention was allocated to the non-verbal input in video than in picture format, and such attention allocation differences were further negatively predicted by learners' PSTM capacity. Additionally, increased attention to the non-verbal element, whether video or picture, resulted in better overall comprehension and larger vocabulary gains in meaning recognition and recall. Our findings give new insights into the role of attention and how it can be maximized, with both theoretical and pedagogical implications for multimodal L2 learning.

Keywords: *Attention, Multimodality, Vocabulary, Comprehension*

Language(s) Learned in This Study: *French, English*

APA Citation: Zhang, P., & Zhang, S. (2025). Attention and learning in L2 multimodality: A webcam-based eye-tracking study. *Language Learning & Technology*, *29*(1), 1–27. https://hdl.handle.net/10125/73626

Introduction

Multimodal input, which integrates different forms of verbal and non-verbal elements such as texts, images, audio, and video, has been increasingly recognized for its potential to enhance both second language (L2) comprehension (Pellicer-Sánchez et al., 2020) and vocabulary learning (Peters, 2019). In the UK, although there has been a growth in provision of L2 multimedia (on platforms like Netflix), anecdotal evidence suggests that it is sparsely used inside the classroom and outside for independent viewing, perhaps because of limited instruction time and learners' relatively low levels of L2 proficiency. Establishing how multimodal input might be most effectively and efficiently used even for lower proficiency learners is therefore a worthwhile endeavor.

The beneficial impact of multimodal input on learning is argued to come from how it enables the activation of multiple sensory channels (Mayer, 2009), hence providing rich contextual and semantic information to facilitate processing. Empirical studies have so far (e.g., Montero-Perez et al., 2014), however, mainly focused on evaluating learning outcomes from multimodal input. Very limited research has explored how L2 learners, especially young learners (Pellicer-Sánchez et al., 2020), make sense of such input by, for example, tracking their attention allocation (AL) using real-time behavioral data (e.g., eye-movements).

*** Corresponding Author:** Pengchong Zhang, anthony.zhang@reading.ac.uk

One further area needing exploration is whether vocabulary learning from and comprehension of multimodal input are predicted by individuals' AL differences. Studies exploring vocabulary learning (Puimège et al., 2023; Wang & Pellicer-Sánchez, 2022) have found that learning gains were positively associated with the amount of AL to the target vocabulary within the verbal element of the multimodal input, commonly in video captions, i.e., written texts representing all audio elements, including L2 dialogue, sound effects, and music for accessibility. No study so far, to our knowledge, has investigated how vocabulary learning is affected by AL to the non-verbal element of the input. Additionally, although some emerging empirical evidence has shown that AL differences in multimodality can predict reading comprehension (Pellicer-Sánchez et al., 2020), that has been found to be not entirely the case for listening comprehension (Suvorov, 2015), and no empirical evidence has been gathered for viewing comprehension. Such contradictory findings necessitate more studies in this area.

The current study, hence, was designed with the following objectives. First, to investigate how young L2 learners' attention is affected by the presence of different types of multimodal input. Second, to explore whether vocabulary learning and viewing comprehension can be explained by individual AL differences. Understanding these issues is not only vital for establishing pedagogical guidelines for using multimodality but also for bringing novel perspectives to existing theories of multimedia learning, taking into account individual differences.

Individual Differences in Attending to Multimodal Input

Individual learners may differ in how they allocate attention to different components of multimodal input, as they have varying levels of sensitivity to non-verbal cues (Gardner, 1999). The additional barrier posed by the L2 may further accentuate these differences. Existing studies have mainly evidenced these AL differences by analyzing learners' AL using eye-tracking to collect real-time eye-movement data (Conklin & Pellicer-Sánchez, 2016). Suvorov (2015) was among the first to use this technology to examine multimodal AL in 33 adult ESL learners. Participants watched six videos: three content-related, featuring lecturers using visual aids (i.e., PowerPoint slides), and three context-related, where the camera focused on lecturers' gestures and facial expressions without slides. The study found that learners allocated significantly more attention to content-related videos, suggesting they found visual aids more useful than non-verbal speaker cues. This aligns with Wagner (2010), who reported that L2 listeners spent less than half their time watching speaker videos during video-based L2 listening tests. A subsequent study by Batty (2021) further found that adult L2 listeners predominantly focused on speakers' facial expressions when taking video-mediated L2 listening tests, often neglecting their gestures.

In the three studies above, multimodal input primarily featured non-verbal visual elements, whereas verbal elements were aural. Although Suvorov (2015) included limited written text on slides in content-related videos, it was redundant with the audio. Warren et al. (2018) further examined AL in multimodal input combining verbal (written) and non-verbal elements. Adult ESL learners ($N = 52$) read a news story with eight pseudowords under three conditions: a bolded pseudoword with a written definition (verbal only), a content-related picture (non-verbal only), or both (multimodal). Learners paid more attention to pictures in the non-verbal only condition than in the multimodal condition, suggesting that multimodality may have caused split attention. Although split attention may lessen focus on individual elements, multimodality has also been shown to enhance overall comprehension and memory by integrating multiple modes of input (Mayer, 2009). Understanding the balance between potential drawbacks like split attention and the broader benefits of multimodal input hence remains critical.

How young learners rather than adults process L2 multimodality has, we believe, only been explored in two studies: Pellicer-Sánchez et al. (2020) and Serrano and Pellicer-Sánchez (2022). In both, elementary school English as a foreign language (EFL) learners watched multimodal materials including written texts accompanied by content-related pictures under two conditions: reading-only vs. reading-while-listening. The reading-while-listening condition prompted learners to pay significantly more attention to the pictures than the reading-only condition, suggesting that the availability of audio freed up visual attention,

allowing learners to seek additional facilitative information from the images. This may have helped them integrate verbal and non-verbal information more effectively. Neither study, however, considered learners' working memory (WM) capacity, which may be a key confounding variable theoretically, pedagogically, and empirically (Teng, 2025). According to the Cognitive Theory of Multimedia Learning (Mayer, 2009), presenting information through both verbal and non-verbal channels reduces WM demands and enhances input uptake by activating dual coding (Paivio, 1986). As verbal and non-verbal information are processed in separate memory systems, this prevents cognitive overload. Given the significant variability in WM capacity among school learners (Kormos & Smith, 2023), processing behaviors may differ accordingly. Further investigation is therefore needed into how WM capacity influences the AL of multimodality among school-aged L2 learners.

Attention Allocation Differences and Comprehension

Multimodal input has been shown empirically to enhance listening comprehension. In reading comprehension, however, a quite different picture was obtained, with no strong evidence supporting its advantage over text-only input (Zhang & Zou, 2022). A key gap in existing research is the lack of consideration for learners' individual AL differences, which may explain the absence of multimodality effects. That was indeed one of the major findings of Pellicer-Sánchez et al. (2020) discussed earlier. In their study, longer attention allocated to the written text was related to less successful reading comprehension, yet greater attention paid to the content-related pictures resulted in better overall reading comprehension. This suggests that multimodal input benefits learners who focus on non-verbal elements but is less effective for those who primarily attend to verbal input.

Further support comes from Gass et al. (2019) and Suvorov (2015). Gass et al. (2019) found that ESL learners with higher WM capacity spent less time watching video captions (verbal element) and achieved better comprehension, suggesting that greater attention to non-verbal elements may have aided understanding. Suvorov (2015) reported a weak correlation between L2 listening comprehension and attention to speakers in context-related videos. Unlike Pellicer-Sánchez et al. (2020), however, Suvorov (2015) found no meaningful link between AL and comprehension when videos included content-related visual aids. Different again, Wang and Pellicer-Sánchez (2023) found no correlation between adult EFL learners' comprehension and attention to L1 or L2 subtitles, that is, written text for spoken dialogues, suggesting that comprehension is not necessarily explained by AL of verbal elements in multimodal input. However, as their study did not measure AL for non-verbal input, its role remains unclear. Overall, these findings highlight the potential importance of AL differences in non-verbal elements for L2 comprehension, yet more evidence is needed given the contradictory findings of Pellicer-Sánchez et al. (2020) and Suvorov (2015) regarding content-related pictures.

Attention Allocation Differences and Vocabulary Learning

An increasing number of studies (e.g., Choi, 2023; Wang & Pellicer-Sánchez, 2022) have examined the relationship between individuals' AL differences in multimodality and L2 vocabulary gains. Most have found a positive correlation between attention allocated to target vocabulary in captions or subtitles and vocabulary acquisition. This is unsurprising, as longer attention on a target word increases the likelihood of noticing it (Schmidt, 1990) and consciously registering it into memory.

Wang and Pellicer-Sánchez (2022) found that although EFL learners spent more time reading L1 translations in subtitles, vocabulary gains were predicted by attention to L2 words in captions or bilingual subtitles, indicating that deeper processing facilitated learning. Similarly, Puimège et al. (2023) reported that total reading time of multiword units in captioned videos significantly predicted learning.

By contrast, Montero-Perez et al. (2015) found that attention to target words in captions only correlated with learning when participants expected a vocabulary test, highlighting the role of learning intentionality. Warren et al. (2018) similarly observed a positive correlation between AL and vocabulary

gains when multimodal input was combined with written and picture glosses. Interestingly, however, the largest gains occurred in the picture-only condition, which elicited the least AL to the pseudowords. This suggests that the effects of different types of multimodal input may outweigh the predictive power of AL differences, if any.

To summarize, several studies have examined the predictive role of AL differences in multimodality on vocabulary learning, but they have only measured attention allocated to target vocabulary within verbal input. To our knowledge, no research has yet investigated whether the AL to the non-verbal element also predicts vocabulary learning. This is crucial, as the non-verbal element is theoretically considered "redundant" when it replicates information from the verbal element (Mayer, 2009). Such redundancy effects have been widely observed in learning through multimedia among first language users (Mayer, 2009). Ample empirical evidence assessing L2 learners' comprehension and learning through multimodality, however, has shown the opposite, namely that such redundancy helps learning (e.g., Peters, 2019). Moreover, research on non-verbal communication suggests that individuals tend to prioritize visual information over other forms of non-visual input (Noller, 1985). This video primacy, or "visual bias" as referred to by Burgoon et al. (2022, p. 508), suggests that when verbal and non-verbal visual elements align, redundancy enables individuals to use visual cues to help interpret verbal information. When they mismatch, however, individuals are more likely to focus on non-verbal visual cues and disregard the verbal information. Finally, given that the type of multimodal input might be a stronger predictor than AL differences (Warren et al., 2018), it is worth further investigating whether the interaction between the type of input and AL differences influences the extent to which learners gain vocabulary from the input.

The Current Study

The current study extends from a previous study (Zhang & Zhang, 2024), which investigated the effects of different types of multimodal input on vocabulary learning. In that study, 43 young English learners of French watched three sets of multimodal input, each representing one input condition: Written+Audio, Written+Audio+Picture, and Written+Speaker+Video. They also completed a vocabulary size test, two target vocabulary tests, a comprehension test, and a phonological short-term memory (PSTM) test. Findings showed that, overall, input with additional non-verbal elements, i.e., demonstrative pictures (static pictures designed to illustrate the meaning of target words) or speaker videos, led to greater learning gains in both form recognition and meaning recall. Additionally, larger gains were generated for the more demanding test of meaning recall under the Written+Speaker+Video condition than under the Written+Audio+Picture condition. Finally, learners' comprehension of the input was the most important moderator of how well they benefited from the input. Those with better comprehension made larger gains than those with poorer comprehension especially when videos were presented.

Going one step further, the current study advances understanding of the role of multimodal input in L2 learning from two perspectives. First, it generates new empirical data on the AL of multimodality by closely examining young learners' AL within two input conditions where non-verbal elements were presented. Second, it investigates whether individual AL differences predict how much new vocabulary is learnt from the input and how well the input is comprehended, through the following research questions:

Research Questions and Hypotheses

RQ1: To what extent does young language learners' attention allocation (dwell time, number of fixations, fixation duration) differ between the two types of multimodal input (Written+Audio+Picture vs. Written+Speaker+Video)?

RQ2: How does attention allocation affect learners' comprehension of the multimodal input?

RQ3: How is vocabulary learning from the multimodal input affected by attention allocation?

We expected learners to allocate more attention to verbal elements overall but hypothesized that the relative attention to non-verbal elements would differ between the two input conditions. Furthermore, individual differences in WM capacity were anticipated to influence AL. Regarding comprehension and vocabulary learning, we hypothesized that greater attention to non-verbal input would enhance both outcomes, with AL further moderating the differences in vocabulary gains between the input conditions.

Method

Participants

Participants were 40 English learners of French (aged 11 to 12) from a range of secondary schools in England who participated outside of school hours. This was a subsample of the 55 learners in the larger study (excluding data from 15 learners who either did not complete the whole experiment or did not have valid eye-movement data captured). Informed consent was first obtained from learners' parents; assent was also gathered from individual learners before they started the experiment. All participants spoke English as their first language, and none spoke or used French outside of school, nor studied formally any other language, as ascertained through a language background questionnaire, adapted from Sabourin et al. (2016). They were considered as basic users (A1-A2, CEFR), having just commenced secondary school with minimal and variable input at primary school (Graham et al., 2017). Learners who completed both experiment sessions received a £20 e-voucher as compensation for their participation.

Design and data collection procedures

Data collection was conducted using the online experiment platform Labvanced (https://www.labvanced.com/, Finger et al., 2017), which featured built-in online eye-tracking technology (see Online Webcam-based Eye-tracking section). The experiment included two sessions: a 30-minute pre-test session, and a combined language learning and post-test session lasting 60–70 minutes, with a two-week gap between them to minimize the pre-test's impact on the learning outcomes of the second session (see Figure 1). Both between- and within-participant designs were employed to boost statistical power. In the pre-test session, participants completed the language background questionnaire, a French vocabulary size test, a vocabulary pre-test, and a PSTM test (see Materials). During the language learning and post-test session, participants first watched three sets of multimodal materials (each in a different input condition – see below), then took a vocabulary post-test and a comprehension test (see Materials). While they were viewing the materials, the built-in camera on the laptop they used automatically tracked their real-time eye-movements. The sequence of test items was randomized for each participant to avoid order effects. All research instruments were piloted before use. Specifically, two target words from one video clip were replaced by more difficult ones as they appeared to be known by most of the participants in the pilot phase.

A short (2-3 minutes) French film clip with bilingual English-French subtitles and six PowerPoint slides, each teaching explicitly a target French word featuring in the film clip, were included in each set of the multimodal materials. One experienced French teacher was responsible for delivering the instruction under three conditions: Written+Audio, Written+Audio+Picture, and Written+Speaker+Video. In the Written+Audio condition, participants saw slides giving the original sentence from the film clip containing the target word, its English translation, the target word with part of speech and English meaning, and an additional example sentence plus English translation (Figure 2). The French teacher read this information aloud. In the Written+Audio+Picture condition, this verbal-only input was supplemented with a picture representing the target word's meaning (Figure 3), and in the Written+Speaker+Video condition with a video of the French teacher reading the slides and adding gestures and facial expressions (Figure 4). Each participant experienced all three conditions with different items. The order of conditions and sets of multimodal input was counterbalanced among participants following a Latin Square design, a feature facilitated by Labvanced. The focus here is solely on the Written+Audio+Picture and Written+Speaker+Video conditions in order to examine the AL of multimodal input where non-verbal elements were included.

Figure 1

Study Design

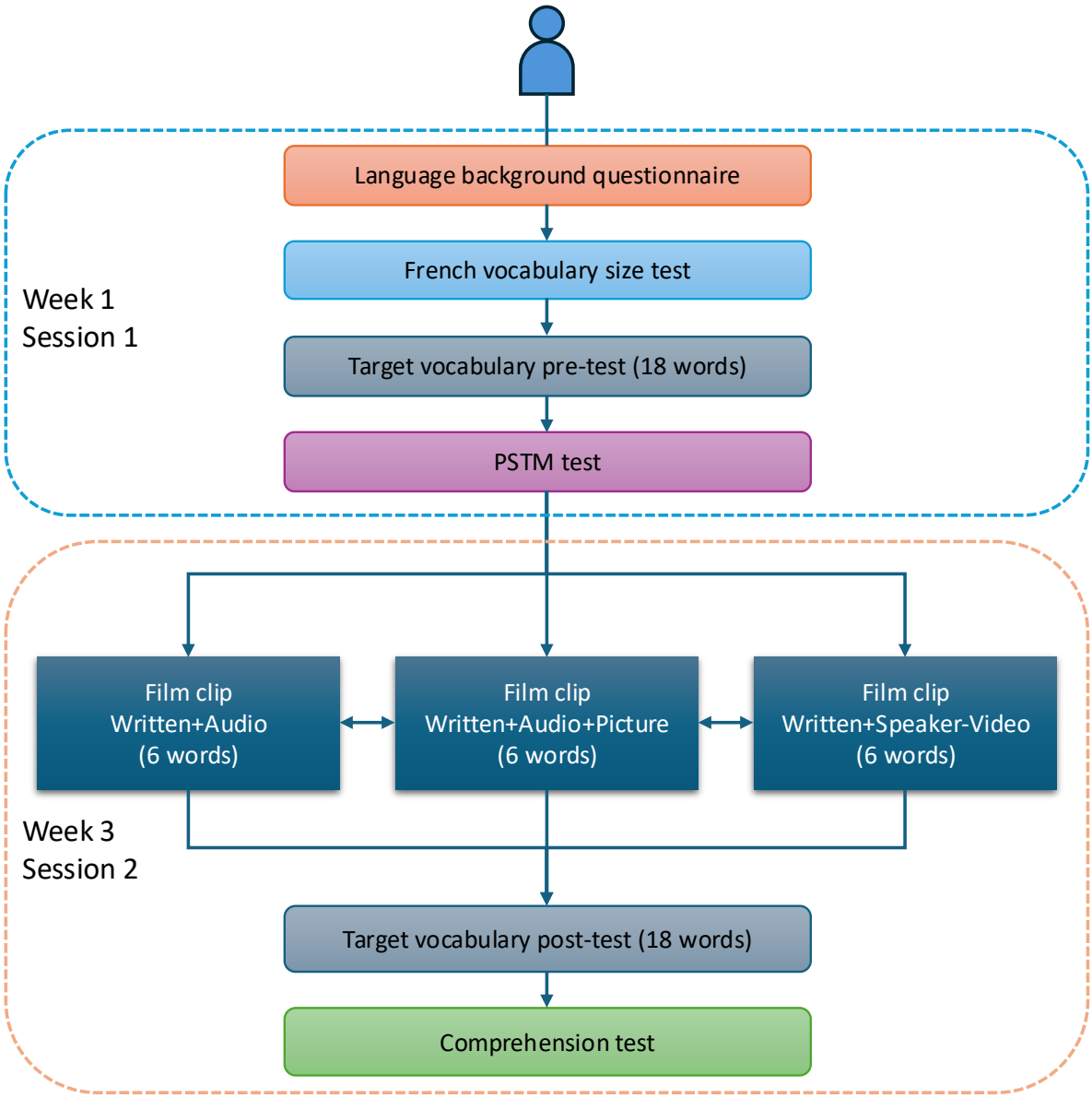


Figure 2

Written+Audio Condition

Il est endormi! *(He is asleep!)*

Endormi

[adjective] ***asleep***

- Voici mon grand-père endormi devant la télévision. (*Here is my grandfather asleep in front of the TV.*)

Figure 3

Written+Audio+Picture Condition

Il est endormi! *(He is asleep!)*

Endormi

[adjective] ***asleep***

- Voici mon grand-père endormi devant la télévision. (*Here is my grandfather asleep in front of the TV.*)

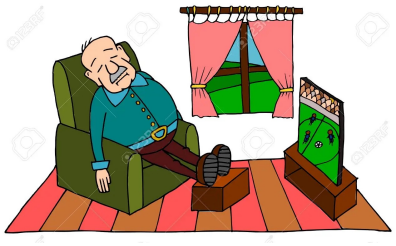


Figure 4

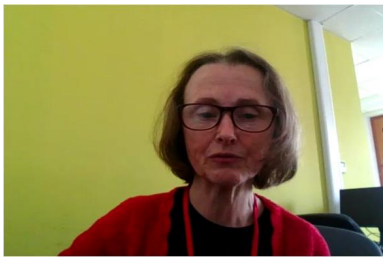
Written+Speaker+Video Condition

Il est endormi! *(He is asleep!)*

Endormi

[adjective] ***asleep***

- Voici mon grand-père endormi devant la télévision. (*Here is my grandfather asleep in front of the TV.*)



Materials

Online Webcam-based Eye-tracking

We used Labvanced webcam-based eye-tracking so that we could collect eye-movement data from participants remotely using their own devices. This webcam-based method allowed eye-tracking with a maximum sampling frequency of 30Hz, although the actual sampling frequencies might vary depending on the specifications of the participants' devices. Studies (e.g., Kaduk et al., 2024) have shown that the eye-movement data acquired with Labvanced were highly correlated (> 80%) with those recorded with the EyeLink 1000 system in lab settings, with an overall accuracy of 1.4° and a precision of 1.1°. Therefore, we were able to maintain the quality of the data collected while also increasing the overall ecological validity of the study design, as most participants would normally view multimodal input on a personal device rather than in a lab with an external eye-tracker.

Before starting the experiment, each learner completed a 55-point, 4-pose calibration and a 3-point recalibration during the learning session after the video clip viewing. The minimum performance of the webcam-based eye-tracking was set to medium-low (> 5*Hz*) as recommended by Labvanced (see https://www.labvanced.com/content/learn/en/guide/eyetracking/) to reach a balance between data quality and hardware requirements. The mean sampling frequency, as calculated after data collection, was 17.7*Hz* (*SD* = 4.88, Min = 7.61, Max = 26.4).

X-Lex

We used a shortened adapted version of the original X-Lex test (Meara, 1992) to measure learners' French vocabulary size (i.e., testing sample range reduced from 10,000 to 5,000 words, and test items reduced from 240 to 120). This version is a Yes/No test assessing form recognition of 120 words, including 100 real French words (20 randomly selected from each of the first five 1000-word frequency

bands) and 20 pseudowords. Learners selected “Yes” for words they knew or could use, and “No” for words they did not know or were invented. Fifty points were awarded for each real word marked “Yes”, with a penalty of 250 points for each pseudoword marked “Yes”. The highest possible score was 5000. Used successfully in large projects with a similar population (Graham et al., 2024), the test demonstrated high reliability, α = .96, 95% *CI* [.94, .97].

Phonological Short-Term Memory Capacity

Learners’ PSTM capacity was measured using a backward digit-span task. Although both the backward and forward digit-span tasks are suitable for non-adults (St Clair Thompson & Allen, 2013), the latter is less demanding, which might have resulted in ceiling effects with no discernible differences among learners. In the selected task, learners heard a sequence of digits with a 1-second interval between each digit and were asked to recall them in reverse order by entering them into a textbox. They completed three 3-digit practice trials before progressing to the formal test, which involved digit sequences of four to seven digits, with three trials per sequence length. All sequences used a first language English speaker recording. Learners received one point for correctly recalling each sequence (Max = 12). The test showed good reliability, α = .85, 95% *CI* [.74, .90].

French Film Clips and Target Words

The three film clips were taken from “Ratatouille” and “Harry Potter and the Philosopher's Stone”. They were selected from the Online Language Learning for All site which hosts materials deemed appropriate for young language learners (Woore et al., 2020). Each clip (2-3 minutes) featured both English and French subtitles and had been scrutinized by an experienced French teacher familiar with England’s foreign languages curriculum and learners’ proficiency level. They then selected 18 target words from the clips (six per clip). All words were from the first two 1000-word frequency bands, as determined by MultilingProfiler (Finlayson et al., 2022).

Comprehension Test

Five multiple-choice questions were designed for each film clip (15 questions in total). Each question required learners to identify one correct answer from three options. The reliability for the test was moderate, α = .77, 95% *CI* [.57, .87].

Vocabulary Pre-Test and Post-Test

The vocabulary pre-test and post-test assessed learners' knowledge of the target words. Both tests, adapted from Montero-Perez et al. (2014), evaluated three aspects of vocabulary knowledge: form recognition (Yes/No test), meaning recall (translate the word into English), and meaning recognition (four-option multiple-choice questions with one correct answer and three distractors), in that order (see Figure 5 for examples). The reliability (see Appendix A) was good for all measurements except for meaning recognition, which was rated as acceptable to moderate, likely due to guessing in the multiple-choice format.

Figure 5

Example Item for Vocabulary Pre-test and Post-test

Form recognition
Have you seen "Endormi" before? (For the post-test, this was phrase as *Has "Endormi" been used in the clips?*)
- Yes
- No

Meaning recall
Translate the word into English
Endormi = _____

Meaning recognition
Choose the correct translation:
Endormi
- Aware
- Asleep
- Sensible
- Unconscious

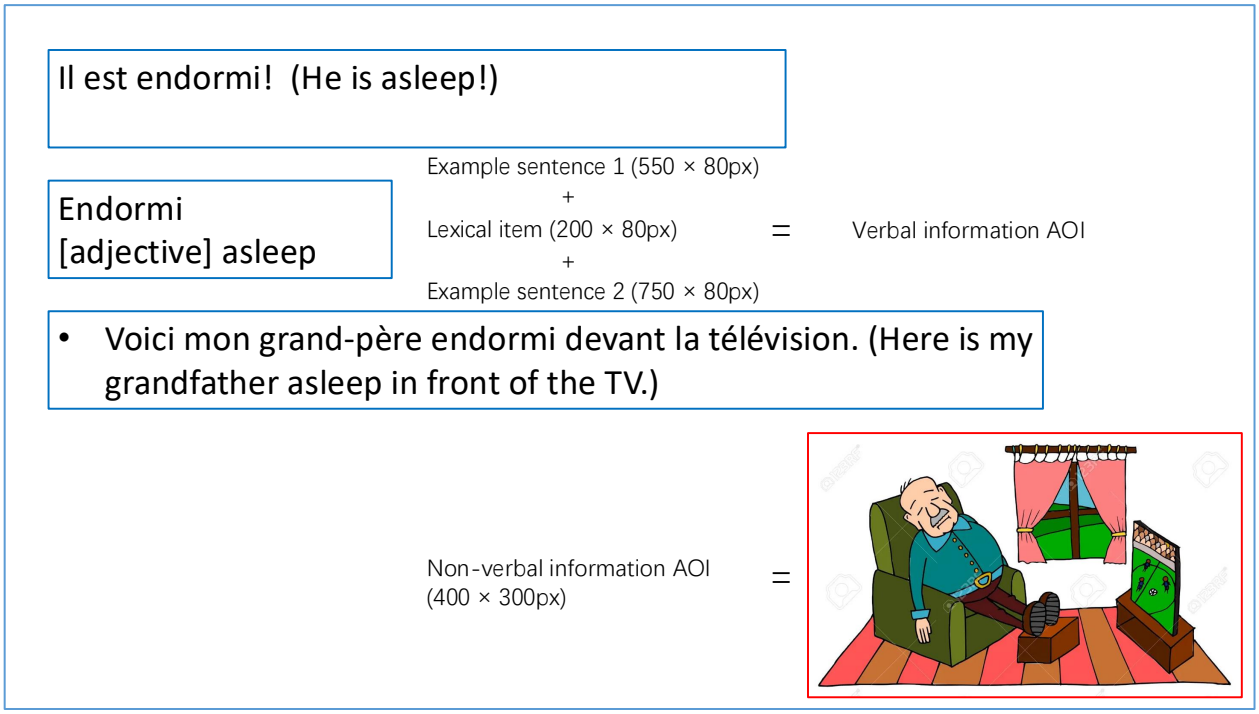
Data Analysis

Eye-movement Data Treatment

We created two Areas of Interest (AOIs): a verbal information AOIs and a non-verbal information AOIs (see Figure 6). The verbal information AOIs included the areas on the slide displaying the lexical item and the example sentences, whereas the non-verbal information AOIs included the areas containing the demonstrative pictures and speaker videos. The sizes and display positions of both AOIs were kept consistent across trials, with each AOI occupying a total of 120,000 pixels. For each AOI, we obtained the average dwell time (i.e., the total amount of time a participant fixates within an AOI), average number of fixations, average fixation duration, as these metrics indicated how much attention was allocated to the AOIs during the learning task. To understand how the AL to verbal and non-verbal information differed under different input conditions, we divided the dwell time/number of fixations/fixation duration for the non-verbal AOIs by those measures for the verbal AOIs and calculated the non-verbal to verbal (NV2V) ratios for these three types of eye-movement data. Similar to the Dwell Time % variable used by Pellicer-Sánchez et al. (2020), the NV2V ratios indicated the attention allocated to the non-verbal elements relative to the verbal elements: the larger such a ratio was, the greater amount of attention was given to the non-verbal input. The use of NV2V ratios could also reduce the complexity of our statistical models without compromising the ability to capture important interactions. That is, we would not get a difficult-to-interpret four-way interaction (e.g., Dwell Time % × AOI × Condition × Time) if AL differences interacted with condition and time and affected learning outcomes. Meanwhile, when the form of non-verbal input (i.e., Picture vs. Video) did affect vocabulary learning, this effect would be captured by the interaction between NV2V ratios and input conditions.

Figure 6

AoI Example



Bayesian Mixed Effects Modeling

We adopted Bayesian mixed-effects models for data analysis, implemented in R (R Development Core Team, 2024) using the brms package (Bürkner, 2021). Bayesian statistics provide a robust framework for modeling by allowing the incorporation of prior knowledge (i.e., informed expectations about the likely effects of the study before collecting new data), which is particularly useful for handling the multimodal and hierarchical data that the current study obtained. This approach can reduce the influence of extreme outliers and enable more nuanced interpretations.

Unlike traditional frequentist statistics, which rely on p-values to determine the significance of an effect, Bayesian statistics focus on estimating the size and credibility of effects. Hence, the key interpretive metric is the 95% credible interval (*CrI*). A 95% *CrI* represents the range within which the true value of the parameter is likely to fall with 95% probability, given the data and the model. If the *CrI* does not include zero (or one when the outcome measure is binary), this suggests that the effect is credibly different from zero. Bayesian models also provide posterior probabilities that quantify the likelihood of a parameter taking on specific values, offering richer information about uncertainty compared to binary decisions based on p-values.

To address RQ1, three models were constructed, one for dwell time NV2V ratio, one for fixation duration NV2V ratio, and one for fixation count NV2V ratio. All models included Condition (Written+Audio+Picture vs. Written+Speaker+Video) as a fixed factor. PSTM was also added to the models as a theory driven fixed factor, as learners' AL is highly associated with their WM capacity. In addition, we further included Condition × PSTM interactions to explore how differences in AL under the two conditions were predicted by WM capacity (Baddeley, 2012; Paivio, 1986). The random effects for all three models included by-participant random intercepts.

Turning to RQ2, the fixed factors for the model included nine continuous predictors: three eye-movement measurements; Xlex; PSTM; and four meaning-related vocabulary measurements (pre-meaning recognition, post-meaning recognition, pre-meaning recall, and post-meaning recall); and one categorical predictor, Condition. The vocabulary measurements were included as fixed factors because of the known strong relationship between vocabulary knowledge and reading/listening comprehension (Zhang & Zhang, 2022). Additionally, interactions between each of the nine continuous fixed factors and Condition were added to the fixed effects structure, exploring whether comprehension differences between the two conditions were predicted by any of the continuous fixed factors. By-Participant and by-Item random intercepts were included to further control the random effects at subject and test item levels.

Finally, three models were built to answer RQ3, one for form recognition, one for meaning recognition, and one for meaning recall. For all models, the initial fixed effects structure included eight fixed factors. There were two categorical factors: Time (pre-test vs. post-test), Condition (Written+Audio+Picture vs. Written+Speaker+Video); and six continuous factors: Comprehension, Xlex, PSTM, and three eye-movement measurements. Additional three-way interactions between each of three eye-movement measurements and the two categorical factors were added to the fixed effects structure to explore how these behavioral variables moderated vocabulary gains between the two input conditions. The random effects structure included both by-item and by-participant random intercepts. By-Item random slopes for Time and by-Participant random slopes for Time were also included to control for the fact that each item was measured twice, and each participant was tested twice, once at the pre-test and once at the post-test. For all Bayesian models, the choice of priors, procedures for model selection, and relevant R code are provided in Appendix B.

Results

Data supporting the results reported in this paper are openly available in Zhang (2025). Descriptive statistics were first calculated by input condition (Written+Audio+Picture vs. Written+Speaker+Video) for all non-behavioral measurements (Appendix C) and behavioral measurements (Appendix D) respectively. Non-behavioral measurements included French vocabulary size, PSTM, and form recognition, meaning recall, and meaning recognition of the target words. Behavioral measurements consisted of three eye-movement measurements (i.e., dwell time, fixation duration, and fixation count) for each of the following: verbal AOIs and non-verbal AOIs as well as NV2V ratios.

Attention Allocation

Model results for fixation duration indicated that neither the fixed factors nor their interactions showed meaningful effects, as they were systematically removed from the fixed effects structure during the cross-validation process. For the model for dwell time, there was an effect of Condition $_{\text{Written+Speaker+Video-Written+Audio+Picture}}$ (β = 0.02, 95% *CrI* [0.01, 0.03]). This suggested that the dwell time NV2V ratios were larger when the Condition was Written+Speaker+Video than when it was Written+Audio+Picture, meaning learners paid more attention to the non-verbal element when the non-verbal element was speaker videos than when it was demonstrative pictures. Finally, the results (Table 1) for the fixation count model indicated that there was an effect of Condition × PSTM interaction (Figure 7). The AL to the different components of the input did not seem to differ between learners with different levels of WM capacity when the condition was Written+Audio+Picture. Within the Written+Speaker+Video condition, however, with every unit of decrease of learners' PSTM capacity, a greater amount of attention was allocated to the non-verbal components than to the verbal components. That is, learners with less PSTM capacity spent more time attending to the speaker videos than those with greater PSTM capacity. In addition, similar to the model for dwell time, there was a simple main effect of Condition, suggesting that when learners' PSTM was centered at the mean, there was a higher fixation count NV2V ratio for the Written+Speaker+Video condition than for the Written+Audio+Picture condition. That is, videos attracted more attention than pictures.

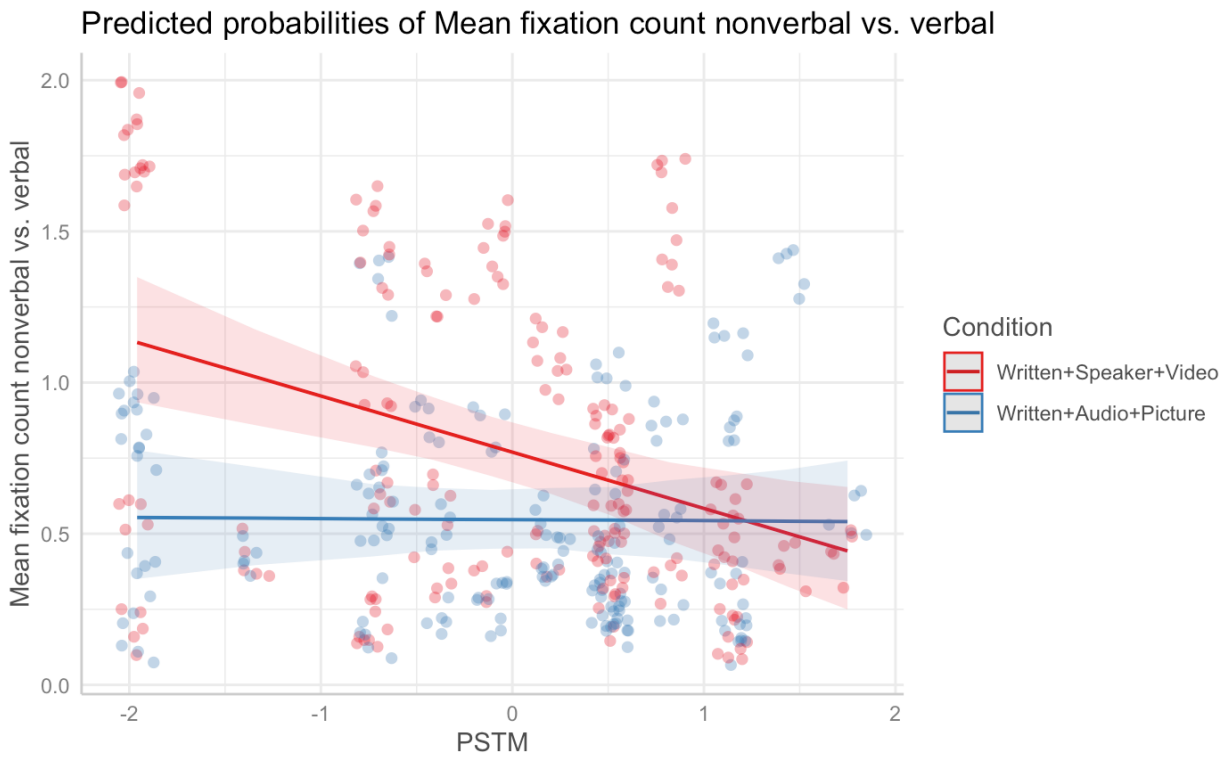
Table 1

Final Model for Fixation Count

	Fixation count NV2V ratio	
Predictors	**Estimates**	**95% *CrI***
Intercept	0.77	0.67 – 0.86
Condition $_{\text{Written+Audio+Picture–Written+Speaker+Video}}$	-0.22	-0.29 – -0.16
PSTM	-0.19	-0.28 – -0.10
Condition $_{\text{Written+Audio+Picture–Written+Speaker+Video}}$ × PSTM	0.18	0.12 – 0.24
$R^2_{Marginal}$ / $R^2_{Conditional}$	0.159 / 0.505	

Figure 7

Effect Plot for the Condition x PSTM Interaction



Comprehension

Table 2 shows the final model results for comprehension. All the interaction terms were removed during model simplification, meaning that differences in comprehension between the two input conditions were not predicted by any of the continuous fixed factors. In addition, only three fixed factors were retained in the final model: Condition, Post-meaning recall, and Dwell time NV2V ratio. The strongest predictor was Post-meaning recall. With one unit increase in target vocabulary knowledge assessed through the meaning recall post-test, learners were 4.68 times more likely to correctly answer the comprehension questions. This was followed by Dwell time NV2V ratio. When it increased by one unit, meaning longer attention

was paid to the non-verbal element, learners were 3.35 times more likely to achieve better comprehension. Finally, learners were 3.10 times more likely to correctly answer the comprehension questions when they were given pictures than when videos were presented.

Table 2

Model Results for Comprehension

	Comprehension	
Predictors	**Odds Ratios (*ORs*)**	**95% *CrI***
Intercept	6.29	2.52 – 18.23
Post-meaning recall	4.68	1.73 – 14.14
Dwell time NV2V ratio	3.35	1.28 – 9.91
Condition $_{\text{Written+Speaker+Video–Written+Audio+Picture}}$	3.10	1.61 – 6.56
$R^2_{Marginal}$ / $R^2_{Conditional}$	0.070 / 0.342	

Vocabulary learning

We first examined the model for form recognition ($R^2_{Marginal}$ = 0.29, $R^2_{Conditional}$ = 0.43). The final retained model included two fixed factors: Time (*OR* = 13.86, *d* = 1.45, 95% *CrI* [6.31, 33.68]) and Xlex (*OR* = 4.57, 95% *CrI* [2.00, 11.09]). These indicated that learners made large pre-post gains in recognizing the form of the target words. The effect size was large, i.e., *d* > 1.40 (Plonsky & Oswald, 2014). Regardless of the test time point, form recognition was positively associated with Xlex scores. With every unit increase in Xlex, the odds ratio was 4.57 times higher for recognizing the form of the target words. The categorical fixed factor, Condition, was removed from the model simplification suggesting that learning gains did not differ between the two input conditions. Form recognition of the target words also did not seem to be predicted by Comprehension, PSTM or any of the eye-movement predictors.

The final model for meaning recognition (Table 3) included the fixed factors of Time, Dwell time NV2V ratio, and Xlex. There were also Time x Dwell time NV2V ratio interactions (Figure 8). These findings suggested that improvement in meaning recognition was positively predicted by learners' AL. With every unit increase in dwell time NV2V ratio, that is more attention allocated to the non-verbal element of the input, participants were 2.50 times more likely to correctly recognize the meaning of the target words. In addition, learners' pre-existing vocabulary knowledge (measured by Xlex) positively predicted their meaning recognition at both time points. The effect size of the simple main effect of Time (*OR* = 12.38, *d* = 1.39) highlighted that the pre-post meaning recognition gains were close to large. Moreover, Comprehension, Condition, PSTM and the other two eye-movement measurements were dropped from the process of model simplification, meaning that they did not add any additional explanations to learners' meaning recognition of the target words.

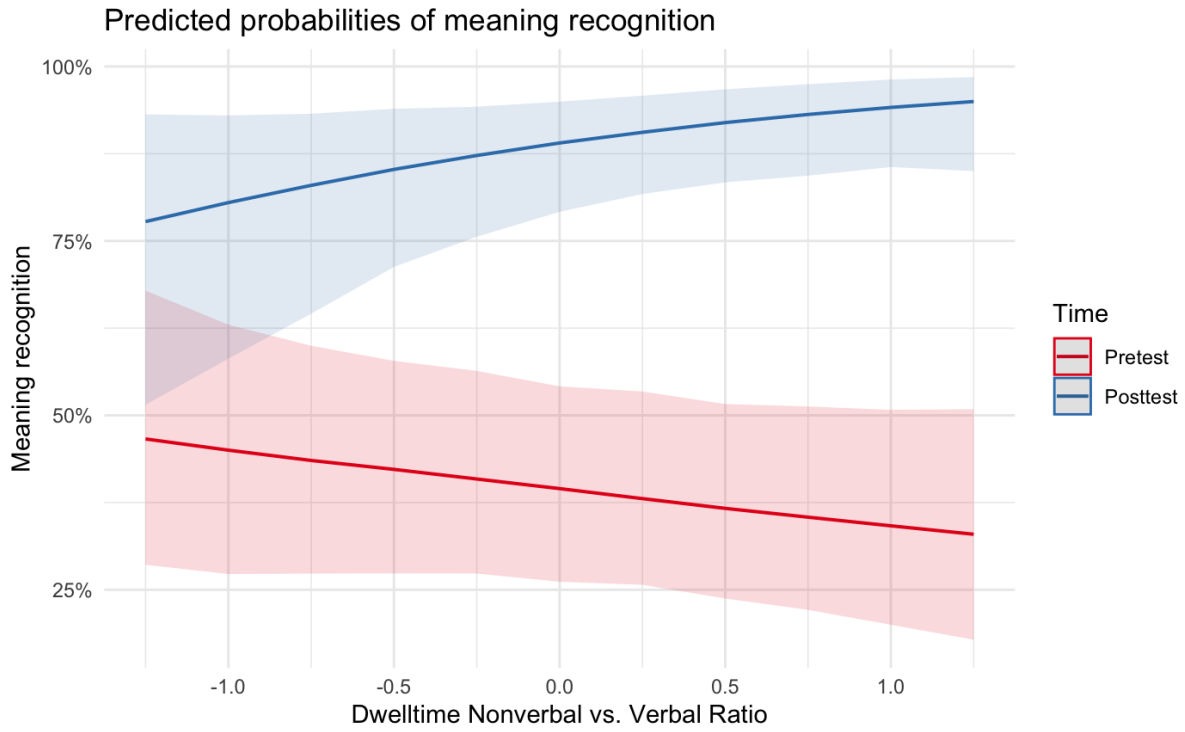
Table 3

Model Results for Meaning Recognition

	Meaning recognition	
Predictors	**Odds Ratios (*ORs*)**	**95% *CrI***
Intercept	0.63	0.34 – 1.14
Time $_{\text{Post-test–Pre-test}}$	12.38	6.92 – 25.40
Xlex	3.09	1.88 – 5.14
Dwell time NV2V ratio	0.79	0.50 – 1.23
Dwell time NV2V ratio × Time $_{\text{Post-test–Pre-test}}$	2.50	1.10 – 5.99
$R^2_{Marginal}$ / $R^2_{Conditional}$	0.272 / 0.373	

Figure 8

Effect Plot for the Time × Comprehension Interaction – Meaning Recognition



A very similar picture was obtained for the final model for meaning recall (Table 4). Closely aligning with the findings for meaning recognition, we found a Time x Dwell time NV2V ratio interaction (Figure 9), indicating that the meaning recall gains from the pre-test to the post-test were positively predicted by dwell time NV2V ratio. With every unit increase in dwell time NV2V ratio, learners were 4.02 times more likely to successfully recall the meaning of a target word at the post-test than at the pre-test. That is, the more attention allocated to the non-verbal element of the input, either demonstrative pictures or speaker videos, the larger gains in target vocabulary knowledge were. Additionally, learners' pre-existing vocabulary size predicted how well they performed in the two target vocabulary tests.

Finally, learners made overall very large meaning recall gains (*OR* = 33.16, *d* = 1.93), evidenced by the effect size of the simple main effect of Time. Again, Condition, PSTM, Comprehension as well as two other eye-movement measurements were removed during model selection, indicating that they did not explain meaning recall.

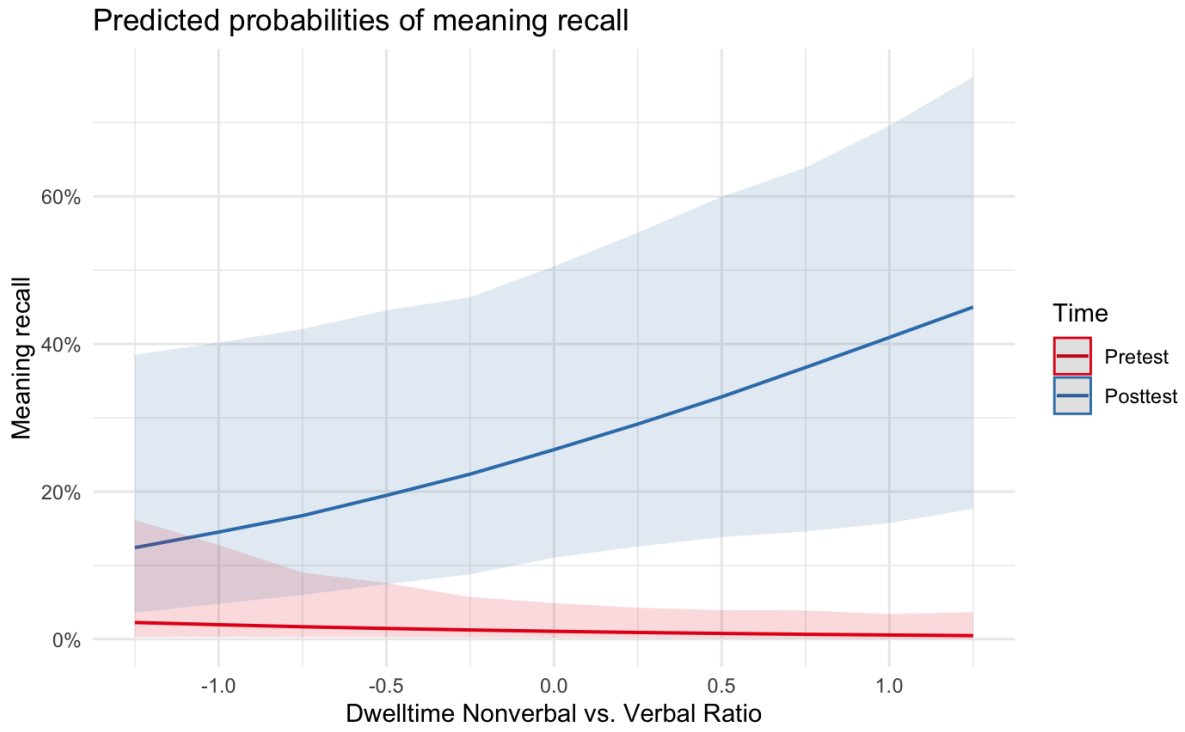
Table 4

Model Results for Meaning Recall

	Meaning recall	
Predictors	**Odds Ratios (*ORs*)**	**95% *CrI***
Intercept	0.01	0.00 – 0.04
Time $_{\text{Post-test–Pre-test}}$	33.16	9.02 – 166.38
Xlex	14.75	5.66 – 40.10
Dwell time NV2V ratio	0.51	0.15 – 1.54
Dwell time NV2V ratio × Time $_{\text{Post-test–Pre-test}}$	4.02	1.14 – 16.78
$R^2_{Marginal}$ / $R^2_{Conditional}$	0.258 / 0.516	

Figure 9

Effect Plot for the Time × Comprehension Interaction – Meaning Recall



Discussion

The current study had a dual focus. First, using an online webcam-based eye-tracking technology, we were able to provide novel empirical data to unpack, for the first time, the complex AL differences of multimodal input among young L2 beginners. Second, we sought to examine whether comprehension and vocabulary learning could be further explained by AL differences in multimodality, in addition to existing theoretically driven individual factors.

Individual differences in AL

In terms of individual AL differences, our findings first revealed that regardless of the type of non-verbal input, learners generally paid more attention to the verbal element than to the non-verbal element. This echoes Warren et al.'s (2018) findings, whereby a significantly larger amount of attention was allocated to the pictures when they were presented alone than when they were accompanied by textual definitions for target vocabulary. Combining verbal and different forms of non-verbal elements in the current study potentially caused split attention (Boers et al., 2017). As the verbal input involved explicit vocabulary instruction, it is not surprising that more attention was allocated to that type of input. Additionally, we found that more attention was paid to the non-verbal element of the multimodal input when it was videos rather than pictures. Significantly more fixations and longer dwell times were observed for the former than for the latter suggesting that participants engaged more with videos, likely perceiving them as more informative or relevant. This finding contradicts Suvorov (2015), who found that videos illustrating the speakers received less attention than those illustrating the content through PowerPoint slides. This discrepancy likely stems from differences in study design and task demands. In Suvorov (2015), content videos included redundant texts on slides, whereas context videos only presented verbal information. As participants were assessed on listening comprehension with detailed content-based questions, they may have prioritized verbal input and minimized visual distractions when watching speakers rather than content-related slides, a pattern also noted in Wagner (2010). In contrast, our study used video clips for the Written+Speaker+Video condition and static pictures for the Written+Audio+Picture condition. Although the pictures were content-related, they did not directly aid vocabulary tests. The speaker videos, however, may have been more beneficial for learning lexical items by providing additional aural-based verbal input and hence attracted more attention.

Our study went one step further than previous studies by examining how learners' PSTM capacity predicted their AL, finding an interesting interaction between input condition and PSTM capacity. Although the amount of attention allocated to the pictures was similar across learners with different levels of PSTM capacity, more attention was allocated to the speaker videos by learners with smaller PSTM capacity than those with larger PSTM capacity. Given the redundancy between the verbal information (text) and speaker videos, lower PSTM capacity learners may have relied more on the speaker videos, where the content aligned with the text, as a preferred input modality to reinforce verbal information through non-verbal visual aids. They therefore chose to spend more time looking at videos than pictures. In contrast, larger PSTM capacity might have allowed others to better regulate and control their AL, regardless of the type of non-verbal input received. Hence, for them, the AL was similar between the two conditions.

AL differences and comprehension

Examining how AL differences predicted comprehension of the input, our findings suggested that when other individual difference factors were controlled for, AL still made a unique contribution to explaining comprehension. Learners who attended more to the non-verbal input, evidenced in the dwell time NV2V ratio, had overall better comprehension. The fact that we did not find a meaningful interaction between eye-movement measurements and input condition indicated that both types of non-verbal input showed similar positive effects on comprehension. This finding further supports existing empirical evidence from Pellicer-Sánchez et al. (2020) and Suvorov (2015) whereby attention allocated to the content-related pictures and context-related videos was found to predict comprehension. More attention allocated to the

non-verbal elements potentially maximized the amount of input being processed at a given time frame (Mayer, 2009) and hence helped learners gain better comprehension. The present study, however, differs from previous studies in that it also found post-meaning recall and input condition predicted comprehension in addition to AL. The predictive power of post-meaning recall was slightly larger than that of AL. This aligns well with evidence from studies examining factors affecting reading/listening comprehension in that vocabulary knowledge is an important factor influencing reading/listening comprehension (Zhang & Zhang, 2022). Finally, comprehension was better when the input condition was Written+Audio+Picture than when it was Written+Speaker+Video. Demonstrative pictures might have strengthened the understanding of the overall meaning of the input as each picture representing the meaning of a target vocabulary appeared in the input. Speaker videos, on the other hand, did not convey any specific meaning-related information and therefore did not help as much as demonstrative pictures for comprehension.

AL differences and vocabulary learning

Regarding how AL affected vocabulary learning, our results revealed that, when other theoretically supported (Teng, 2025; Zhang & Zhang, 2024) individual difference factors (Xlex, PSTM, Comprehension) were controlled for, dwell time NV2V ratio positively predicted the learning gains for meaning recognition and recall of the target words. Learners who attended more to the non-verbal elements of the input benefited more in vocabulary learning. This extends from previous studies (Puimège et al., 2023; Wang & Pellicer-Sánchez, 2022) examining the relationship between AL and vocabulary learning, providing additional evidence for the positive effects of attention allocated to the non-verbal visual input on vocabulary learning. It also aligns with the non-verbal communication literature (Burgoon et al., 2022; Noller, 1985), which suggests that when the same information is presented in both visual and non-visual formats, learners rely on visual input to interpret the non-visual input. Additionally, such evidence more directly indicates that the redundancy principle (Mayer, 2009), cannot be simply applied to the L2 learning context as it is applied to multimedia learning in the L1. Although the pictures and videos duplicated the verbal element of the input, it did not seem to be redundant or to cause cognitive overload as learners who paid more attention to made larger vocabulary gains.

It is worth noting, however, that this predictive effect of AL was found to be larger for meaning recall than for meaning recognition. For form recognition, no predictive effects were detected. These findings indicate that more attention allocated to non-verbal input seems to be especially useful in facilitating the acquisition of more demanding vocabulary knowledge (i.e., meaning recall) and may not be equally beneficial for less demanding vocabulary knowledge (i.e., form recognition) (González-Fernández & Schmitt, 2020). This is potentially because the intentional vocabulary learning approach adopted by the study might have been particularly useful in stimulating learners' noticing (Schmidt, 1990) of the target words. That higher level of noticing particularly advantaged the learning of less demanding knowledge and outweighed the small impact of attention differences. Finally, there was no meaningful interaction between AL and input condition, suggesting more words were learnt by those who paid more attention to the non-verbal elements regardless of the type of input. This finding contradicts our previous study (Zhang & Zhang, 2024), where the Written+Speaker+Video condition demonstrated larger vocabulary gains than the Written+Audio+Picture condition when both were compared to the Written+Audio condition. This discrepancy is not unexpected, however, as AL was not considered in the earlier study. The absence of meaningful interactions between AL and input condition in the current study highlights the importance of taking finer measurements, such as AL differences, into consideration when modeling how multimodal input affects vocabulary gains.

Limitations and Future Research

Our study is among the first to use webcam-based eye-tracking technology to collect behavioral eye-movement data from young learners. Although such technology achieves accuracy comparable to traditional lab-based eye-trackers (Kaduk et al., 2024), it partially depends on the processing speed of

participants' devices, learners' home laptops in this case, introducing potential environmental variability. Consequently, we did not measure attention to the smallest AOIs, such as specific target vocabulary. Future studies could mitigate this by providing uniform, high-speed devices to participants, enabling more precise data collection for smaller AOIs. Despite these constraints, allowing participants to use their own devices enhanced ecological validity and yielded data that better reflect natural processing compared to highly controlled lab settings.

Another limitation involves the design of the form recognition test. The test instructions differed slightly between the pre-test and post-test to align with the study's aims, potentially imposing different processing demands. Additionally, considering the high testing burden, we did not add French pseudowords to reduce the risk of learners guessing dishonestly. Future research could use consistent instructions across test phases and incorporate additional pseudowords to provide more comprehensive data while balancing testing burden.

Conclusions and Implications

The study established its novelty by adopting webcam-based eye-tracking technology to examine young L2 learners' AL of different types of multimodal input and how AL differences explained comprehension and vocabulary learning of the input. Our findings highlighted that the two input conditions triggered different AL patterns. A larger amount of attention was allocated to the non-verbal input when it was speaker video than when it was picture, suggesting learners spent more time engaging with videos. Such AL differences were further predicted by learners' WM capacity. Learners with smaller WM capacity spent significantly more time attending to videos than those with larger WM capacity. AL, however, did not differ among learners with different levels of WM capacity regarding pictures. Theoretically, these findings provide crucial evidence for the role of WM capacity in dual coding (Paivio, 1986). Learners with larger WM capacity are more likely to regulate their attention across different components of multimodal input, yet those with limited WM may focus more on the element carrying greater informational load, potentially leading to cognitive overload. Given that this study was conducted with young learners at the early stages of L2 learning, the findings suggest that multimodal input should be designed to support learners with varying cognitive capacities. Rather than requiring teachers to assess WM directly, a practical approach would be to balance verbal and non-verbal input, ensuring that neither element is overly complex. Using clear, concise language and aligning visual cues closely with verbal content can help young learners process information more effectively.

We further found that both successful comprehension and larger vocabulary gains were positively correlated to the amount of attention allocated to the non-verbal input. At a theoretical level, this feeds into the existing empirical evidence supporting the fact that the redundancy principle for multimedia learning (Mayer, 2009) needs to be reconsidered within the L2 learning context. Duplicated information presented in different formats, processed through different memory channels, may not be truly redundant for L2 learners. In our study, those who paid more attention to those "redundant" elements made larger vocabulary gains and had better comprehension. Within the foreign language classroom, pedagogical input can be designed in such a way that key knowledge is usefully presented in more than one single format, e.g., both verbal and non-verbal, to facilitate learning and comprehension.

Acknowledgements

This study was supported by University of Reading Research Fellowship (Improving Foreign Language in the UK: Transforming Classroom Language Teaching through Multimedia; A368700).

References

Baddeley, A. (2012). WM: Theories, models, and controversies. *Annual Review of Psychology, 63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Batty, A. O. (2021). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing, 38*(4), 511–535. https://doi.org/10.1177/0265532220951504

Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System, 66*, 113–129. https://doi.org/10.1016/j.system.2017.03.017

Burgoon, J.K., Manusov, V., & Guerrero, L.K. (2021). *Nonverbal communication* (2nd ed.). Routledge. https://doi.org/10.4324/9781003095552

Bürkner, P. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software, 100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Choi, S. (2023). Visual saliency in captioned digital videos and learning of English collocations: An eye-tracking study. *Language Learning & Technology, 27*(1), 1–21. https://hdl.handle.net/10125/73536

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research, 32*(3), 453–467. https://doi.org/10.1177/0267658316637401

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. In *International Conference on Computational Social Science*. https://www.labvanced.com/static/2017_IC2S2_LabVanced.pdf

Finlayson, N., Marsden, E., & Anthony, L. (2022). *MultilingProfiler (Version 3)* [Computer software]. University of York. https://www.multilingprofiler.net/

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. Basic Books.

Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A working-memory, eye-tracking study. *Language Learning & Technology, 23*(2), 84–104. https://doi.org/10125/44684

Gelman, A. (2024, July). *Prior choice recommendations*. GitHub. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2017). Early language learning: The impact of teaching and teacher factors. *Language Learning, 67*(4), https://dx.doi.org/10.1111/lang.12251

Graham, S., Zhang, P., Hofweber, J., Fisher, L., & Krüsemann, H. (2024). Literature and second language vocabulary learning: The role of text type and teaching approach. *The Modern Language Journal, 108*(3), 579–600. https://doi.org/10.1111/modl.12946

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Kaduk, T., Goeke, C., Finger, H., & König, P. (2024). Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the EyeLink 1000. *Behavior Research Methods*, *56*, 5002–5022. https://doi.org/10.3758/s13428-023-02237-8

Kormos, J., & Smith, A. M. (2023). *Teaching language to students with specific learning differences*. Multilingual Matters.

Lüdecke et al., (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60), Article 3139. https://doi.org/10.21105/joss.03139

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. https://doi.org/10.1016/j.jml.2017.01.001

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Meara, P. (1992). *EFL vocabulary tests*. CALS University of Wales Swansea.

Montero-Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned Video: An eye‐tracking study. *The Modern Language Journal, 99*(2), 308–328. https://doi.org/10.1111/modl.12215

Montero-Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology, 18*(1), 118–141. https://doi.org/10125/44357

Noller, P. (1985). Video primacy: A further look. *Journal of Nonverbal Behavior, 9*(1), 28–47. https://doi.org/10.1007/BF00987557

Paivio, A (1986). *Mental representations: A dual coding approach*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195066661.001.0001

Pellicer-Sánchez, A., Tragant, E., Conklin, K., Rodgers, M., Serrano, R., & Llanes, Á. (2020). Young learners' processing of multimodal input and its impact on reading comprehension: An eye-tracking study. *Studies in Second Language Acquisition, 42*(3), 577–598. https://doi.org/10.1017/S0272263120000091

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly, 53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Puimège, E., Montero Perez, M., & Peters, E. (2023). Promoting L2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research, 39*(2), 471–492. https://doi.org/10.1177/02676583211049741

R Development Core Team. (2024). *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing. https://www.R-project.org

Sabourin, L., Leclerc, J.-C., Lapierre, M., Burkholder, M. & Brien, C. (2016). The language background questionnaires in L2 research: Teasing apart the variables. In L. Hracs (ed.), *Proceedings of the 2016 Annual Conference of the Canadian Linguistics Association* (pp. 1–15). https://cla-acl.ca/pdfs/actes-2016/Sabourin_etal_CLA2016_proceedings.pdf

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Serrano, R., & Pellicer-Sánchez, A. (2022). Young L2 learners' online processing of information in a graded reader during reading-only and reading-while-listening conditions: A study of eye-movements. *Applied Linguistics Review, 13*(1), 49–70. https://doi.org/10.1515/applirev-2018-0102

Sivula, T., Magnusson, M., & Vehtari, A. (2020). *Uncertainty in Bayesian leave-one-out cross-validation based model comparison*. arXiv. https://arxiv.org/abs/2008.10296

St Clair-Thompson, H., & Allen, R. J. (2013). Are forward and backward recall the same? A dual-task study of digit recall. *Memory & Cognition, 41*(4), 519–532. https://doi.org/10.3758/s13421-012-0277-2

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 463–483. https://doi.org/10.1177/0265532214562099

Teng, M. F. (2025). Effectiveness of captioned videos for incidental vocabulary learning and retention: the role of working memory. *Computer Assisted Language Learning*, *38*(1–2), 206–234. https://doi.org/10.1080/09588221.2023.2173613

Wagner, E. (2010). Test-takers' interaction with an L2 video listening test. *System, 38*(2), 280–291. https://doi.org/10.1016/j.system.2010.01.003

Wang, A., & Pellicer-Sánchez, A. (2023). Examining the effectiveness of bilingual subtitles for comprehension: An eye-tracking study. *Studies in Second Language Acquisition, 45*(4), 882–905. https://doi.org/10.1017/S0272263122000493

Wang, A., & Pellicer-Sánchez, A. (2022). Incidental vocabulary learning from bilingual subtitled viewing: An eye‐tracking study. *Language Learning, 72*(3), 765–805. https://doi.org/10.1111/lang.12495

Warren, P., Boers, F., Grimshaw, G., & Siyanova-Chanturia, A. (2018). The effect of gloss type on learners' intake of new words during reading: Evidence from eye-tracking. *Studies in Second Language Acquisition, 40*(4), 883–906. https://doi.org/10.1017/S0272263118000177

Woore, R. Graham, S., & Arndt, H. L. (2020). *Online language learning for all (OLLA)*. PDC in MFL. https://pdcinmfl.com/online-language-learning-for-all-olla/

Zhang, R., & Zou, D. (2022). Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning, 35*(4), 696–742. https://doi.org/10.1080/09588221.2020.1744666

Zhang, P. (2025). *Vocabulary learning through multimodal input*. [Dataset]. University of Reading. https://doi.org/10.17864/1947.001343

Zhang, P., & Zhang, S. (2024). Multimedia enhanced vocabulary learning: The role of input condition and learner-related factors. *System, 122*, Article 103275. https://doi.org/10.1016/j.system.2024.103275

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research, 26*(4), 696–725. https://doi.org/10.1177/1362168820913998

Appendix A. Reliability for Vocabulary Pre-test and Post-test

Tests time point	Form recognition	Meaning recall	Meaning recognition
Pre-test	α = .88, 95% *CI* [.78, .92]	α = .89, 95% *CI* [.80, .92]	α = .68, 95% *CI* [.56, .79]
Post-test	α = .87, 95% *CI* [.77, .91]	α = .90, 95% *CI* [.84, .93]	α = .76, 95% *CI* [.60, .84]

Appendix B. Choice of Priors, Procedures for Model Selection, and R Code

Choice of Priors

For RQ1, as the outcome variables were on a ratio scale, Bayesian linear mixed effects models were performed. Generic weakly informative priors (Gelman, 2024) were used whereby a normal distribution was set for all predictor variables with a mean of 0 and *SD* of 1. For RQ2 and 3 where the outcome variables were on a binary scale, Bayesian generalized linear mixed effects models were chosen. Weakly informative priors followed the recommendations of Gelman et al. (2008) and Gelman (2024). All nonbinary predictors were scaled to have a mean of 0 and *SD* of 0.50. For all predictor terms, a student's t distribution with four degrees of freedom was then set. Regarding the intercept term, the distribution was scaled to have a mean of 0 and *SD* of 10 whereas the distribution for all other predictor terms was centered at the mean and had a *SD* of 2.50.

Procedures for Model Selection

As the structure of some of our models (for RQ2 and 3) was rather complex, although theoretically driven, backward model simplification was undertaken through cross-validation. Such an approach is believed to be able to maintain the Type-I error rate as close as possible to the one of the full model yet substantially increase the overall power (Matuschek et al., 2017). Every step of model simplification involved a comparison between the simplified model and the original model. For each comparison, an elpd_diff (the expected log pointwise predictive density difference) value was calculated. In cases where an elpd_diff value was smaller than 4, the simplified model was judged to have a better fit and therefore was retained before continuing with further model simplification (Sivula et al., 2020). For RQ1, meaningful effects were judged when the value zero did not fall into the range of the 95% *CrI*. For RQ2 and 3, however, meaningful effects were confirmed when the value of one did not cross the range of the 95% *CrI*.

As the models for RQ2 and RQ3 included multiple eye-movement measurements as fixed factors, there was a risk that these measurements were highly correlated and hence had issues of multicollinearity. Although we believe such issues could be addressed by performing cross-validation, in order to further ensure that our final models did not have multicollinearity issues, we obtained VIF (variance inflation factor) for each final model using the check_collinearity function within the performance package (Lüdecke et al., 2021). All final models had VIF scores below five indicating no significant issues of multicollinearity (James et al., 2023).

R Code

```
priors.weak <- c(prior(student_t(4, 0, 10), class=Intercept), prior(student_t(4, 0, 2.5), class=b), prior(student_t(4, 0, 2.5), class=sd))

priors.weak2 <- c(prior(normal(0, 1), class=Intercept), prior(normal(0, 1), class=b), prior(normal(0, 1), class=sd))

model_fixation_count<- brm(mean_fixnumber_nv2v ~ Condition * WM + (1 | Subject), prior = priors.weak2, data = df_eyetracking[which(df_eyetracking$Task_Name == "comprehension_test" & df_eyetracking$Condition != "Written+Audio"), ])

model_fixation_duration <- brm(mean_fixtime_nv2v ~ Condition * WM + (1 | Subject), prior = priors.weak2, data = df_eyetracking[which(df_eyetracking$Task_Name == "comprehension_test" & df_eyetracking$Condition != "Written+Audio"), ])

model_dwell_time <- brm(mean_dwelltime_nv2v ~ Condition + (1 | Subject), prior = priors.weak2, data = df_eyetracking[which(df_eyetracking$Task_Name == "comprehension_test" & df_eyetracking$Condition != "control"), ])
```

```
model_comprehension_nv2v <- brm(answer_clip_comprehension ~ centered_mean_dwelltime_nv2v + Condition + (1|Trial_Id) + (1 |Subject), family =bernoulli, prior = priors.weak, data = df_eyetracking[which(df_eyetracking$Task_Name == "comprehension_test" & df_eyetracking$Condition != "control"), ])

model_form_recognition <- brm(answer_word_used ~ Xlex + Time + (1 + Time |Trial_Id) + (1 + Time |Subject), family =bernoulli, prior = priors.weak, data = df_eyetracking[which(df_eyetracking$Task_Name == "vocab_pretest2" & df_eyetracking$Condition != "control" | df_eyetracking$Task_Name == "vocab_test" & df_eyetracking$Condition != "control"), ])

model_meaning_recognition <- brm(answer_word_recognition ~ Xlex + centered_mean_dwelltime_nv2v * Time + (1 + Time |Trial_Id) + (1 + Time |Subject), family =bernoulli, prior = priors.weak, data = df_eyetracking[which(df_eyetracking$Task_Name == "vocab_pretest2" & df_eyetracking$Condition != "control" | df_eyetracking$Task_Name == "vocab_test" & df_eyetracking$Condition != "Written+Audio"), ])

model_meaning_recall <- brm(answer_word_translation ~ Xlex + centered_mean_dwelltime_nv2v * Time + (1 + Time |Trial_Id) + (1 + Time |Subject), family =bernoulli, prior = priors.weak, data = df_eyetracking[which(df_eyetracking$Task_Name == "vocab_pretest2" & df_eyetracking$Condition != "control" | df_eyetracking$Task_Name == "vocab_test" & df_eyetracking$Condition != "Written+Audio"), ])
```

Appendix C. Descriptive Statistics for Non-behavioral Measurements

Variables	Condition	*M*	*SD*	Min	Max
Xlex		498.81	542.07	0.00	2450.00
PSTM		0.54	0.50	0.00	1.00
Comprehension	Written+Audio+Picture	0.88	0.33	0.20	1.00
	Written+Speaker+Video	0.78	0.41	0.20	1.00
Form recognition – Pre-test	Written+Audio+Picture	0.26	0.44	0.00	0.83
	Written+Speaker+Video	0.28	0.45	0.00	1.00
Form recognition – Post-test	Written+Audio+Picture	0.71	0.46	0.00	1.00
	Written+Speaker+Video	0.69	0.46	0.00	1.00
Meaning recognition – Pre-test	Written+Audio+Picture	0.42	0.50	0.00	0.83
	Written+Speaker+Video	0.41	0.49	0.00	0.83
Meaning recognition – Post-test	Written+Audio+Picture	0.80	0.40	0.33	1.00
	Written+Speaker+Video	0.82	0.38	0.17	1.00
Meaning recall – Pre-test	Written+Audio+Picture	0.09	0.29	0.00	0.67
	Written+Speaker+Video	0.09	0.29	0.00	0.67
Meaning recall – Post-test	Written+Audio+Picture	0.36	0.48	0.00	0.83
	Written+Speaker+Video	0.35	0.48	0.00	1.00

Appendix D. Descriptive Statistics for Behavioral Measurements (Raw Dwell Time and Fixation Duration Data Are in Milliseconds)

Variables	Condition	*M*	*SD*	Min	Max
Dwell time – Non-verbal	Written+Audio+Picture	828.84	774.05	0.00	3231.98
	Written+Speaker+Video	834.58	796.45	0.00	3486.10
Dwell time - Verbal	Written+Audio+Picture	7790.99	4470.72	872.65	20721.25
	Written+Speaker+Video	7830.41	4349.49	2993.40	28030.16
Dwell time - NV2V ratio	Written+Audio+Picture	0.10	0.08	0.00	0.33
	Written+Speaker+Video	0.12	0.13	0.00	0.63
Fixation duration - Non-verbal	Written+Audio+Picture	222.48	23.34	172.24	293.22
	Written+Speaker+Video	222.97	22.16	143.22	284.67
Fixation duration - Verbal	Written+Audio+Picture	222.88	15.60	197.48	285.22
	Written+Speaker+Video	223.11	12.15	197.38	275.32
Fixation duration - NV2V ratio	Written+Audio+Picture	1.00	0.08	0.81	1.22
	Written+Speaker+Video	1.00	0.09	0.68	1.26
Fixation count - Non-verbal	Written+Audio+Picture	11.33	8.48	2.00	41.00
	Written+Speaker+Video	14.52	12.94	3.00	74.00
Fixation count - Verbal	Written+Audio+Picture	20.08	7.41	4.91	43.42
	Written+Speaker+Video	19.86	8.25	8.25	51.00
Fixation count - NV2V ratio	Written+Audio+Picture	0.54	0.33	0.13	1.37
	Written+Speaker+Video	0.78	0.69	0.09	3.94

About the Authors

Pengchong Zhang is a Lecturer in Second Language Learning at the University of Reading, UK. His research interests include technology-enhanced language learning, school-based language learning, multimodality, vocabulary, and language comprehension. Pengchong Zhang is the corresponding author.

E-mail: anthony.zhang@reading.ac.uk

ORCiD: https://orcid.org/0000-0002-2136-4984

Shi Zhang is an Assistant Professor at College of Foreign Languages and Cultures, Chengdu University of Technology, China. His research interests include interdisciplinary research on bilingualism, language processing, and language acquisition/education.

E-mail: shizhang.chn@gmail.com

ORCiD: https://orcid.org/0000-0002-1459-5215