Text First or Picture First? Evaluating Two Modes of Multimodal Input for EFL Vocabulary Meaning Acquisition

Jia Yu and Xiaobin Liu

Abstract
Several studies and assumptions support the advantages of presenting pictures before texts, such as the Sequencing Effects of Integrated Model of Text and Picture Comprehension (the Sequencing Effects of ITPC Model) and the Scaffolding Assumption. However, few studies have examined the efficacy of text-first or picture-first multimodal input on second language (L2) learners’ vocabulary meaning acquisition. To examine whether the assumptions can be applied to the L2 vocabulary context, this study compares the effect of text-first and picture-first multimodal input on English as a Foreign Language (EFL) learners’ word meaning learning and retention. The multimodal materials in this study included written forms of English words, audio, first language (L1) translations, and pictures. For text-first input, Chinese translations preceded pictures while it was the reverse for picture-first input. Pre-, post-, and delayed post-tests were conducted to examine learners’ gains and retention of word meaning. The results showed that though both groups achieved significant gains, participants learning with picture-first input performed significantly better and reported positive attitudes. This study reveals the advantageous effects of presenting pictures before L1 translations on L2 vocabulary meaning learning, thus indicating that the Sequencing Effects of IPTC Model and the Scaffolding Assumption can be applied to the L2 multimodal vocabulary learning context, and providing pedagogical implications for EFL teachers and material developers.

Keywords
multimodal input, text-first, picture-first, L2 word meaning

Introduction
Vocabulary knowledge plays an indispensable role in second language (L2) performance (O’Toole, 2022; Polakova & Klimova, 2022). According to I. S. P. Nation (2013), what is involved in learning a word includes form, meaning, and use. Among these aspects of word knowledge, many language learners find it particularly difficult to memorize word meaning (Brown, 2010; Chung, 2018; Chung & Fung, 2022; Schmitt, 2008).

In recent years, different modalities have been integrated into vocabulary teaching in the English as a foreign language (EFL) learning environment. Among them, pictures are considered to be beneficial to vocabulary learning, especially when it comes to word meaning (Andrà, 2020; Hasnine et al., 2019; Mayer, 2001; Sydorenko, 2010). Furthermore, first language (L1) translations also make a profound impact on L2 vocabulary learning, and learners often employ L1 translations to obtain word meaning (Alghamdi, 2016; Schmitt, 1997; Swan, 1997; Tavakoli & Gerami, 2013). In addition, verbal information (spoken words) is often used in EFL vocabulary learning and has been found to be useful (Dubois & Vial, 2000; Plass et al., 1998; Rassaei, 2018). Therefore, it is plausible to incorporate these modalities (pictures, L1 translations, and audio), in a process widely known as multimedia learning (I. S. P. Nation, 2013) into the EFL vocabulary learning process to help learners acquire word meaning.

However, how to arrange the modalities’ sequence in the multimodal EFL vocabulary learning material is a question worth discussing. The primary goal of this study is to evaluate the effects of different sequences of multimodal presentations on EFL learners’ gains and retention of English vocabulary meaning. The modalities employed in this study were written L2 words, audio, L1 translations, and pictures. These modalities were arranged and presented in two different sequences, including (a) text-first input: L2 words and L1
translators appeared first and pictures appeared later with audio; (b) picture-first input: L2 words and pictures appeared first and L1 translations appeared later with audio. Moreover, participants' attitudes toward the different inputs were examined. In this way, the present study serves to suggest a beneficial way to organize multimodal materials in the EFL vocabulary learning process.

**Literature Review**

**Multimodal Input on L2 Vocabulary Learning**

Multimodal input refers to the input that integrates distinct semiotic modes, such as language, images, and sound (Kress & van Leeuwen, 2001). Paivio’s (1986) dual coding theory (DCT) suggests that information is processed in two separate channels, verbal and non-verbal. The verbal channel deals with information such as spoken and written words, while the non-verbal channel processes information such as pictures. Mayer (2001) defined multimedia learning as the learning in which students receive information presented in more than one mode, such as in pictures and words, and he proposed the multimedia effect, namely that learners learn better from materials containing texts and pictures than from text-only ones. In his wake, many studies have explored the potential benefits of multimodal input on L2 vocabulary learning by combining L1 translations and pictures (Acha, 2009; Bisson et al., 2014; Emirmustafaoğlu & Gökmen, 2015; Lian et al., 2017; Moradan & Vafaei, 2016). The findings of previous research indicate that both L1 translations and pictures contribute to L2 word learning, and that they might be effective in different ways. For example, Kookjeong (2015) compared the effects of L1 translations and pictures on Korean EFL learners’ comprehension and retention of new English words. The results of the comprehension test indicated that the group presented with L1 translations of new words outperformed the group with pictures. Participants also expressed a preference for L1 translations when it came to comprehending new words. However, the results of the retention test suggested a reverse situation. The group presented with pictures performed better than the group presented with L1 translations did, and participants reported that they favored L1 translations rather than pictures when memorizing words.

**Sequence of Texts and Pictures**

According to Mayer’s (2001) Temporal Contiguity Principle of multimedia learning, students can benefit more from a simultaneous presentation of corresponding words and pictures than from a successive presentation. Most of the previous studies have presented L1 translations and pictures at the same time, and revealed the advantages of adding pictures in L2 vocabulary learning. Nevertheless, simultaneously presenting pictures and L1 translations does not always lead to better performance in L2 vocabulary learning. Boers et al. (2017), for instance, conducted a study to examine whether adding pictures to L1 translation glosses enhanced L2 learners’ vocabulary uptake from reading. The study found that presenting both pictures and L1 translations simultaneously did not help learners to remember the target words any better than providing L1 glosses only. Therefore, how to make use of pictures and L1 translations in L2 vocabulary learning is a question worth exploring.

In the L1 reading context, studies have revealed the advantages of presenting pictures before texts (picture-first materials; e.g., Eitel et al., 2013; Schüler & Mayer, 2020). The possible explanation that lies behind the benefits of presenting pictures before texts is the Sequencing Effects of ITPC Model (Schnitz & Bannert, 2003). According to the Sequencing Effects of ITPC Model, learners can construct mental models through first-presented visual images, and the mental models have a scaffolding function for textual comprehension. Similar assumptions are made by the Scaffolding Assumption (Eitel et al., 2013), which assumes that even a brief processing of pictures before texts could yield a mental scaffold that helps students better process subsequent texts.

The study conducted by Eitel et al. (2013), for instance, showed with the help of eye-tracking that a transient (no more than 2 seconds) presentation of a picture provided a scaffolding function in mental model construction. Therefore, they formulated the hypothetical Scaffolding Assumption to explain the advantage of picture-to-text processing. According to the assumption, a picture is helpful to extract a global spatial structure very quickly through early processing. This global spatial structure remains active in working memory and is then reactivated during succeeding textual processing, thereby enhancing text comprehension.

However, the Sequencing Effects of ITPC Model and the Scaffolding Assumption were put forward in the context of L1 reading. Whether they can be applied to L2 vocabulary learning is unknown. This study is intended to verify the applicability of the assumptions in the L2 vocabulary learning context. To be specific, this study examines whether presenting picture-first input (pictures before L1 translations) has advantageous effects on gaining and recalling L2 word meaning.

Furthermore, most previous studies have investigated the effectiveness of multimodal input on incidental L2 vocabulary acquisition from reading or listening, and these studies mainly focused on vocabulary form (Çakmak & Erçetin, 2018; Pavia et al., 2019; Wang & Lee, 2021). Few studies have compared the effects of text-first multimodal input (L1 translations before pictures) and picture-first multimodal input (pictures before L1 translations) on explicit L2 vocabulary meaning learning and retention.

In addition, verbal audition information has commonly been adopted for delivering instructional content and can explicitly connect the form of a word (through pronunciation) with its meaning. Adopting verbal auditory information
in multimodal vocabulary instruction is standard practice (Kaplan-Rakowski et al., 2019). However, few studies have examined the effect of pronunciation in combination with text-first or picture-first material.

As word meaning is one of the aspects that EFL learners need to know (I. S. P. Nation, 2013), this study combines written L2 words, L1 translations, pictures and audio as multimodal input, and compares the effects of text-first input (L1 translations preceding pictures) and picture-first input (pictures preceding L1 translations) on EFL learners’ vocabulary meaning learning and retention.

**Methodology**

**The Present Research**

In this study, the effects of different sequences of L1 translations and pictures on the gains and retention of EFL learners’ vocabulary meaning were compared. In addition, students’ attitudes toward the different sequences were investigated. The present study was guided by the following research questions:

RQ1: Which sequence of multimodal input contributes to EFL learners’ L2 word meaning learning and retention?

RQ2: What are EFL learners’ attitudes toward different sequences of multimodal vocabulary learning materials?

**Participants**

Sixty-six Grade 8 students whose native language was Chinese in two intact classes in Guangzhou, Guangdong Province, China were engaged in this study. Their age ranged from 14 to 15 years old. Participants had been learning English as a second language for 8 years. None of them reported having any previous learning experiences abroad. They had similar English proficiency (intermediate level), as reported by their head teachers. The two classes were randomly divided into two groups: the text-first group (TG; \(N=31\)) and the picture-first group (PG; \(N=35\)). Information about the two groups is presented in Table 1.

|                  | Male (%) | Female (%) | Total |
|------------------|----------|------------|-------|
| Text-first group | 13 (41.9)| 18 (58.1)  | 31    |
| Picture-first group | 11 (31.4)| 24 (68.6)  | 35    |

**Instruments**

**Pre-test.** To select words unknown or unfamiliar to participants as the target words, a pre-test containing 150 English words was developed by the researchers. The 150 words were randomly selected from the 3,500 EFL words that must be learned by Chinese senior high students to participate in National College Entrance Examination. Taking the form-meaning matching test suggested by P. Nation (1990) as a reference, vocabulary tests in this study adopted multiple-choice questions. Students were required to choose the correct Chinese meaning for each L2 word. There were four choices for each word: one correct answer, two distractors, and one choice named “I don’t know” to avoid students guessing the answer. The correct answer was the same as provided by the 3,500 EFL word list, and the distractors were created following two criteria: (a) the length of the distractors should be equal to that of the correct answer (following Heaton, 1988; Hughes, 2002) and (b) the part of speech of the distractors should be the same as the correct answer’s (Heaton, 1988).

**Post-test/delayed post-tests.** A post-test and a delayed post-test were developed to examine participants’ vocabulary meaning learning and retention. About 38 target words were included, and the items were the same as the items in the pre-test.

**Questionnaire.** To explore participants’ attitudes toward different sequences of multimodal vocabulary learning materials, a questionnaire was adapted from Lin and Yu’s (2017). The questionnaire was intended to get a deeper understanding of participants’ performances on vocabulary tests and their feelings toward the presentation of L1 translations and pictures. The questions concerned participants’ overall perception of multimodal vocabulary learning materials (three items), their evaluation of the effectiveness of L1 translations and pictures (five items), and their satisfaction with the multimodal input (two items). The 10 items were measured in the form of a five-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). A question about participants’ preferred sequence of L1 translations and pictures was added, and they were required to select one sequence and give their reasons. The choices included (a) presenting the two simultaneously, (b) L1 translations followed by pictures, and (c) pictures followed by L1 translations. This question was to better understand participants’ performance during the study. The Cronbach’s alpha reliability was .890, indicating the relevant high reliability of the questionnaire.

**Learning Materials**

Target words were selected according to the results of pre-tests. Those words that more than 80% of the participants could not recognize were chosen as target words (38 words in total). Multimodal vocabulary learning input was created by the researchers. Each word was learned from a video containing the L2 written word, written pronunciation, aural pronunciation, L1 translation, and a picture illustrating the meaning of the L2 word. Each video was presented for 8s (Sheshadri et al., 2020). For the text-first group, L2
written word, written pronunciation, and L1 translation were presented for the first 4s, and then aural pronunciation and a picture were given for the last 4s (Figure 1). For the picture-first group, the sequence of L1 translations and pictures was reversed. L2 written word, written pronunciation, and pictures were presented for the first 4s, and then aural pronunciation and L1 translation were given for the last 4s (Figure 2).

Procedures

The experimental design is presented in Figure 3. The text-first group learned L2 vocabulary through text-first materials with L1 translations preceding pictures, while the picture-first group learned with pictures preceding L1 translations.

One week before the experiment, all participants together with their parents completed the written informed consent forms. After that, participants were required to complete the pre-test carefully, so there was no time limit for the pre-test. The central part of the study was conducted in participants’ English classes with the help of their English teacher. Participants were presented with multimodal videos to learn L2 vocabulary meaning. The videos were played on one big screen positioned in the front of the classroom. The sound was played from speakers which were placed on the right and left sides of the classroom. Participants watched the videos only once. After watching, an immediate post-test was assigned to all participants to evaluate their gains in L2 vocabulary meaning. In addition, to examine students’ attitudes toward the multimodal materials, a questionnaire was given to all students. One week after the experiment, a delayed post-test was assigned to students to check their retention of the meaning of the L2 vocabulary items. The pre- and delayed post-tests as well as the questionnaire were completed on mobile phone on wjx.cn, while immediate post-tests were finished by pen and paper.

Scoring and Analyses

To check participants’ gains and retention of target words’ meaning, they were required to choose the correct L1 translation for each L2 word. Participants received 1 point for each correct answer and 0 points for each wrong answers. As there were 38 target words, the highest possible score was 38. All scoring was done by the first author.

The experimental design was a 2 × 3 mixed model, with Group (CG and PG) as the between-subject factor and Time (pre-test, post-test, and delayed post-test) as the within-subject factor. The analysis was conducted using a two-way repeated-measures ANOVA, with a significance level of .05. The result of Mauchly’s Test of Sphericity was not statistically significant (p = .113 > .05), which indicates that sphericity was not violated.

Results

Participants’ Vocabulary Meaning Gains

Table 2 displays the descriptive statistics for both groups’ performance on all three tests. As it shows, both groups made considerable learning progress after the experiment, but PG achieved higher scores than CG did in both post- and delayed post-tests (see also Figure 4).

Tables 3 and 4 illustrate the results of the tests including within-subjects and between-subjects effects. The results revealed a significant large main effect for time [F (2,128)=320.60, p < .001, η² = .83], which means the changes in participants’ vocabulary pre-test, post-test, and delayed post-test results were significant. A significant large main effect for group was found as well [F (1,64)=15.56, p < .001, η² = .20], indicating a significant difference between the two groups’ performances. In addition, time and group interaction were significant on the participants’ vocabulary tests [F (2,128)=11.01, p < .001, η² = .15], which suggests that changes in participants’ vocabulary tests were different between the groups.
Figure 3. The experimental design.

Table 2. Descriptive Statistics of Both Groups' Performance on the Three Tests.

|                      | Pre-test | Post-test | Delayed post-test |
|----------------------|----------|-----------|-------------------|
|                      | M (SD)   | M (SD)    | M (SD)            |
| Text-first group (N=31) | 4.68 (4.69) | 22.90 (7.31) | 16.90 (7.14) |
| Picture-first group (N=35) | 4.31 (4.12) | 30.77 (4.88) | 22.31 (7.65) |

Figure 4. Mean scores for both groups on the three tests.
The results of pairwise comparisons for the three tests in both groups are presented in Table 5. It indicates that the scores of the post-test were significantly higher than that of the pre-test as well as the delayed post-test in both text-first group (MD_{post-pre} = 18.23, p < .001; MD_{post-delayed} = 6.00, p < .001) and picture-first group (MD_{post-pre} = 26.46, p < .001; MD_{post-delayed} = 8.46, p < .001), and the scores of the delayed post-test were significantly higher than that of the pre-test in text-first group (MD_{delayed-pre} = 12.23, p < .001) as well as picture-first group (MD_{delayed-pre} = 18.00, p < .001).

Table 6 displays the results of pairwise comparisons for the two groups in three tests. As it shows, no significant difference was found between the two groups in pre-test scores (MD_{TG-PG} = 0.36, p = .739). However, a significant difference was found between the two groups in the immediate post-test (MD_{TG-PG} = −7.87, p < .001), which represents a large effect size (Plonsky & Oswald, 2014). A significant difference also existed in the two groups’ delayed post-test scores (MD_{TG-PG} = −5.41, p = .004), a medium effect size according to Plonsky and Oswald (2014). The picture-first group scored significantly higher than the text-first group did in both post-tests and delayed post-tests.

To sum up, the results indicate that the groups showed little difference at the beginning of the experiment, suggesting that both groups had the same level in terms of target vocabulary meaning at the onset of the experimental period. With experimental intervention, though both groups had significant gains in learning target word meaning, the picture-first group achieved significantly better results than the text-first group did in post- and delayed post-tests. It should be noted that although both groups had significant losses in delayed post-test compared with post-test scores, they still had significant gains in delayed post-test compared with pre-test scores.

### Quantitative Results of the Questionnaires

Four participants in TG did not finish the questionnaire, so there were 62 questionnaires in total. The results are presented in Table 7.

As shown in Table 7, a large number of the participants held positive attitudes concerning the multimodal vocabulary learning materials. As is reported, the multimodal input was interesting (50, 80.7%) and motivating (49, 79.0%), and participants were willing to use the materials in the future (47, 75.8%).

Most participants responded that pictures were more attractive (47, 75.8%), and L1 translations were effective in terms of learning (52, 83.9%) and recalling word meaning (48, 77.4%). Even more participants agreed that pictures were beneficial to learning (56, 90.4%) and recalling (54, 87.1%) word meaning.
As for the two groups’ satisfaction with the multimodal input, Table 7 indicates that nearly half of the participants in TG thought that text-first materials helped them to learn word meaning (16, 59.2%). Though many participants were uncertain about the effectiveness of text-first materials on their word meaning retention (11, 40.7%), there were a few participants who agreed or strongly agreed (13, 38.1%). Meanwhile, a great number of participants in PG believed that picture-first materials were helpful in learning (30, 85.7%) and recalling (29, 82.8%) word meaning. No one in PG clearly stated that they disagreed with the effectiveness of picture-first materials.

A question about participants’ preferred sequence of L1 translations and pictures was included. There were three choices: (a) simultaneous input, (b) text-first input, and (c) picture-first input. Participants were also required to write down their reasons. Five answers were excluded because they did not provide useful information. The answers were coded by the first author, and the results are presented in Table 8.

As shown in Table 8, most participants in PG preferred picture-first input (N=25) because it was impressive and could provide scaffolding for them to learn word meaning. Most students in TG preferred simultaneous input (N=11) for its convenience and advantage in offering better understanding, while a few students preferred text-first input (N=8) as it could enhance meaning learning.

### Discussion

This study intends to evaluate the effects of text-first and picture-first multimodal input on L2 word meaning learning and retention. The multimodal input contained written L2 words, audio, L1 translations and pictures. For text-first input, L1 translations preceded pictures for 4s, while for picture-first input, pictures proceeded L1 translations for 4s.

RQ1 addressed the better sequence of multimodal input for EFL learners’ L2 word meaning learning and retention. Based on the vocabulary meaning tests, both TG and PG acquired significant gains with the multimodal input, which
are in line with previous studies indicating the benefits of multimodal input on EFL vocabulary learning (Alzahrani & Roberts, 2021; Khezrlou et al., 2017; Peters, 2019; Ramezanali & Faez, 2019). However, the results seem inconsistent with the findings claimed by some other previous studies, which concluded that students who learned with the multimodal input method did not achieve better vocabulary gains (Boers et al., 2017; Lin & Yu, 2017). Possible reasons lie in the presentation sequence of pictures and texts in the present study and will be further discussed.

**Better Performance of PG in Immediate Post-Test**

Though both groups achieved significant gains immediately after the experiment, the results of the post-test indicated that PG outperformed TG significantly, which is consistent with previous research (Eitel et al., 2013; Merchie et al., 2021; Schüler & Mayer, 2020).

From the perspective of mental construction, the better learning outcomes of PG may be explained by the Sequencing Effects of ITPC Model (Schnitz & Bannert, 2003) and the Scaffolding Assumption (Eitel et al., 2013), both of which suggest that a brief look at pictures first may offer a mental scaffold for learners, and enhance subsequent textual processing. As some participants revealed in the questionnaires, “Picture-first input can give me some idea about the word, so when L1 translations appeared, I can better understand the word meaning.” Presenting pictures first provided students with scaffoldings of word meaning, therefore PG achieved better learning outcomes than CG did.

Similarly, according to Levin et al. (1987), visual displays such as pictures provide learners with a supportive function that helps them to understand and learn difficult materials. The supportive function seems to be obvious regarding learners of lower prior knowledge. In this study, participants had little prior knowledge about the target words’ meaning. In this case, pictures-based input supported participants’ understanding and learning of L1 translations, thus leading to the better learning results of PG. Students in PG reported: “I can get the general idea of the word meaning from looking at the pictures first, and I was very satisfied when the following L1 translations were similar to my interpretation.”

Moreover, working memory has been found to exert an impact on L2 vocabulary acquisition (Cheung, 1996; Ellis, 1996; Martin & Ellis, 2012). Brief processing of a picture first places only low demands on working memory. Therefore, there is enough space left in learners’ working memory for later textual processing (Kulhavy et al., 1985, 1993). The cited studies furnished evidence that pictorial presentation before textual presentation facilitates learning compared with the reversed situation. This effect can be explained by the assumption that a global spatial structure of a picture can be maintained as an intact unit in working memory when reading a subsequent text, and the later textual processing will not exceed working memory capacity. The spatial structure of the picture can offer a mental scaffolding that has advantageous effects on learning from the following text (Gyselinck et al., 2008; Kulhavy et al., 1993). Therefore, picture-first multimodal input was effective in that presenting pictures first not only gave learners mental scaffolds, but also allowed students more working memory capacity for processing the following L1 translations.

Furthermore, according to the Noticing Hypothesis (Schmidt, 1990), attention was vital to transmitting the information. Without enough effective attention, students fail to encode information successfully and transfer it to long-term memory. In this study, more than half of the learners in PG reported that picture-first input was impressive and interesting, saying that “It (picture-first input) could better arouse my attention to the learning materials,” and “pictures first would not make me feel bored when learning vocabulary. It aroused my interest and made me focus on the following materials (L1 translations).” Pictures as visual aids were more vivid, colorful, and attractive, and when they were presented first, they could arouse students’ learning curiosity and interest to some extent. Once involved in the learning environment with enough curiosity and interest, students could improve their learning effect as their concentration improved their short-term memory. Thus, this may be a reason accounting for the better performance of PG in the immediate vocabulary post-test.

**Better Performance of PG in Delayed Post-Test**

The delayed post-test results indicated that PG achieved significantly better scores than TG did, which suggests that participants in PG recalled more words than TG did 1 week after the experiment.

According to DCT, pictorial and textual information are processed in two separate but related memory stores (i.e., nonverbal and verbal memory stores). Extracting information from the nonverbal memory store mechanically stimulates the relevant information in the verbal memory store and vice versa. In this way, if the information from one of the two stores can be retrieved, it is enough to extract the corresponding information from the other store. In consequence, it will be easier to retrieve information from the two memory stores if the two are better connected. Based on DCT, Kulhavy et al. (1993) proposed the model of working memory operations and applied it to the processing of maps. Kulhavy et al. (1993) explained that maps are intactly and holistically stored in memory as a unit, and that the unit can be maintained in working memory as a single chunk (Miller, 1956). Therefore, if a map is presented before the corresponding text, information extracted from the map can be kept as an intact unit in working memory. When learners then encode information from subsequent text, the processing will not exceed their cognitive capacity. This permits simultaneous processing of map (picture) and text information, resulting in
better-connected memory stores and therefore better retrieval in the process of information recall.

In this study, the pictorial information was the same as that provided by L1 translations, which means that these two kinds of information were connected closely. Learners in PG also stated, “I could get a general idea about the word meaning from the picture.” Pictures in the present study played the role of maps used in the study by Kulhavy et al. (1993), giving learners a holistic idea about the meaning of L2 words. Therefore, the model of working memory operations can also explain the better retention results of PG.

In addition, the recency effect may account for PG’s greater retention results since informational retention ought to be better for the resource (text or picture) presented second, that is to say, in closer temporal propinquity to the assessment (Baddeley & Hitch, 1993). In this study, the assessed knowledge was word meaning (L1 translations) rather than pictures. For picture-first input, L1 translations were presented most recently prior to the assessment, which may contribute to PG’s better performance in the delayed post-test.

Participants’ Attitudes Toward Learning Materials

RQ2 addressed EFL learners’ attitudes toward different sequences of multimodal vocabulary learning materials. According to the questionnaire, a great majority of participants reported great satisfaction with the multimodal learning materials, which to some extent can explain both groups’ significant progress in vocabulary learning. Mayer’s multimedia effect suggests that learners’ understanding is improved by the combination of text and pictures, compared to text-only or picture-only. In this study, L1 translations and pictures were combined, so that the multimedia effect may be partly responsible for both groups’ significant gains in L2 word meaning.

Specifically, students in PG showed preference for picture-first input, while most students in TG preferred simultaneous input rather than text-first input. Overall, participants in PG thought that picture-first input was impressive because “Pictures coming before texts impressed me and made me want to know more about the word,” and it could provide scaffolding because “Pictures first helped me to guess (interpret) word meaning, and I could check (the interpretation) when L1 translations occurred. It was like a competition and made the learning process full of fun.”

Though some participants in TG believed that text-first input helped to enhance word meaning learning, more participants thought that simultaneous input was convenient and could provide a better understanding of word meaning. As part of Mayer’s (2001) multimedia learning, the Temporal Contiguity Principle suggests that learners can benefit more from the simultaneous presentation of words and pictures than from the successive presentation of the two. Considering the principle and participants’ responses in the study, simultaneous input can also be beneficial. Learners in TG reported that “Combining pictures and L1 translations together may help to understand the word better” and “It (simultaneous input) is convenient and saves me time.” These answers indicate that in this study, text-first multimodal input was not perceived to be a good choice for participants to learn the English word meaning. Rather, picture-first and simultaneous input are better.

Conclusions, Implications, and Limitations

This study reveals that the Sequencing Effects of ITPC Model and the Scaffolding Assumption can be applied to an EFL multimodal vocabulary learning context. The study provides empirical evidence that when integrating pictures and L1 translations for EFL learners to study new words, it is better to present pictures before L1 translations. Picture-first input can help learners achieve greater gains in L2 vocabulary meaning due to advantages in the deployment of mental construction, working memory, and learners’ attention. Moreover, supported by DCT and recency effects, learners can benefit more from picture-first input in terms of vocabulary meaning retention. Also, learners themselves reported higher satisfaction with picture-first input thanks to its impressiveness and scaffolding function.

The results of the study have a number of important implications. For EFL vocabulary material developers, pictures, L1 translations and audio can be incorporated to explain L2 word meaning, as both methods resulted in significant gains in vocabulary learning. Still, PG achieved significantly higher results in both post- and delayed post-tests. Therefore, it will be more effective to present pictures before L1 translations to help students learn and memorize the word meaning. As some participants replied that picture first attracted them more so they paid more attention to the picture-first input, presenting picture first in multimodal input can help learners focus on the input and acquire better learning results. For L2 teachers, both text-first input and picture-first input can be adopted, and teachers can choose one of them in different contexts. When it comes to teaching new vocabulary, adopting picture-first multimodal input instead of text-first multimodal input may improve students’ learning effects, as the former method showed greater gains in this study. But when the pedagogical aim is to help learners review the vocabulary and consolidate their knowledge, it may be better to adopt text-first input, as several participants preferred text-first input because it enhanced their learning.

Nevertheless, there are still several limitations to this study. First, only the effects of text-first and picture-first input were compared. Further studies can compare the effects of text-first input, picture-first input and simultaneous input on L2 word meaning learning, as some participants voiced their preference for simultaneous input. Second, to better understand participants’ processing of different sequential
inputs, further investigation employing eye-tracking techniques is needed. The pre- and post-test approach reflects only the learning results, but what is more important is the participants’ learning process with different sequential inputs. Eye-tracking technology can better visualize how learners process different sequential inputs and thus provide more evidence for the better learning outcome of text-first input or picture-first input, as the experiment conducted in the study by Eitel et al. (2013).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Center for Language Cognition and Assessment, Guangdong Province, China. It’s also the result of Guangdong 13th Five-Year Plan Project of Philosophy and Social Science (GD20 WZX01-02).

ORCID iD

Xiaobin Liu https://orcid.org/0000-0001-5395-4420

References

Acha, J. (2009). The effectiveness of multimedia programmes in children’s vocabulary learning. British Journal of Educational Technology, 40, 23–31.

Alghamdi, E. A. (2016). Untangling multimedia effects on EFL incidental vocabulary learning via playing an online hidden-object game. International Journal of Computer-Assisted Language Learning and Teaching, 6(1), 24–39.

Alzahrani, S., & Roberts, L. (2021). The effect of visuospatial designing elements of zoomable user interfaces on second language vocabulary acquisition. System, 96, 102396.

Andri, C., Mathias, B., Schwager, A., Macedonia, M., & von Kriegstein, K. (2020). Learning foreign language vocabulary with gestures and pictures enhances vocabulary memory for several months post-learning in eight-year-old school children. Educational Psychology Review, 32(3), 815–850.

Baddeley, A. D., & Hitch, G. (1993). The recency effect: Implicit learning with explicit retrieval? Memory & Cognition, 21, 146–155.

Bisson, M.-J., van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). The role of verbal and pictorial information in multimedia incidental acquisition of foreign language vocabulary. Taylor & Francis, 68(7), 1306–1326.

Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? System, 66, 113–129.

Brown, D. (2010). What aspects of vocabulary knowledge do textbooks give attention to? Language Teaching Research, 15(1), 83–97.

Çakmak, F., & Ergün, G. (2018). Effects of gloss type on text recall and incidental vocabulary learning in mobile-assisted L2 listening. ReCALL, 30(1), 24–47.

Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary language. Developmental Psychology, 32(5), 867–873.

Chung, E. (2018). Revisiting second language vocabulary teaching: Insights from Hong Kong in-service teachers. The Asia-Pacific Education Researcher, 27(6), 499–508.

Chung, E., & Fung, D. (2022). ‘I realised I only knew the word partly’: Student beliefs about vocabulary knowledge and learning in different grades and proficiency levels. The Language Learning Journal. Advance online publication. https://doi.org/10.1080/09571736.2021.2023204

Dubois, M., & Vial, I. (2000). Multimedia design: The effects of relating multimodal information. Journal of Computer Assisted Learning, 16, 157–165.

Eitel, A., Scheitel, K., & Schüler, A. (2013). How inspecting a picture affects processing of text in multimedia learning. Applied Cognitive Psychology, 27, 451–461.

Ellis, N. C. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. Quarterly Journal of Experimental Psychology, 49, 234–250.

Emirnustafaoglu, A., & Gökmen, D. U. (2015). The effects of picture vs. translation mediated instruction on L2 vocabulary learning. Procedia - Social and Behavioral Sciences, 199, 357–362.

Gyselinck, V., Jamet, E., & Dubois, V. (2008). The role of working memory components in multimedia comprehension. Applied Cognitive Psychology, 22, 353–374.

Hasnine, M. N., Flanagan, B., Akcapinar, G., Ogata, H., Mouri, K., & Uosaki, N. (2019, July 26–31). Vocabulary learning support system based on automatic image captioning technology [Conference session]. International Conference on Human-Computer Interaction, Orlando, FL, United States (pp. 346–358). Springer.

Heaton, J. B. (1988). Writing English language tests (Longman handbooks for language teachers) (pp. 54–55). Longman.

Hughes, A. (2002). Testing for language teachers (p. 181). Cambridge University Press.

Kaplan-Rakowski, R., & Loranc-Paszylk, B. (2019). The impact of verbal and nonverbal auditory resources on explicit foreign language vocabulary learning. System, 85, 102–114.

Khezrlou, S., Ellis, R., & Sadeghi, K. (2017). Effects of computer-assisted glosses on EFL learners’ vocabulary acquisition and reading comprehension in three learning conditions. System, 65, 104–116.

Kookjeong, P. (2015). L2 vocabulary learning: Lexical route of L1 translation & conceptual route of visuals. Korean Journal of Applied Linguistics, 31, 95–122.

Kress, G., & van Leeuwen, T. (2001). Multimodal discourse: The mode and media of contemporary communication (pp. 21–22). Arnold.

Kulhavy, R. W., Lee, J. B., & Caterino, L. C. (1985). Conjoint retention of maps and related discourse. Contemporary Educational Psychology, 10, 28–37.

Kulhavy, R. W., Stock, W. A., & Kealy, W. A. (1993). How geographic maps increase recall of instructional text. Educational Technology Research and Development, 41, 47–62.

Levin, J. R., Anglin, G. J., & Carney, R. N. (1987). On empirically validating functions of pictures in prose. The Psychology of Illustration, 1, 51–86.

Lian, Y.-Y., Chen, C.-M., & Li, Y.-R. (2017, July 9–13). Effects of collaborative multimedia annotations on elementary school students’ vocabulary learning performance [Conference
Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting.

Plass, J., Chun, D. M., Mayer, R., & Leutner, D. (1998). Supporting.

Peters, E. (2019). The effect of imagery and on-screen text on for-

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learn-

Mental representations: A dual coding approach

Paivio, A. (1986). Mental representations: A dual coding approach. Oxford University Press.

Moradan, A., & Vafaie, M. (2016). The effect of glosses on inciden-
tal vocabulary learning of Iranian EFL learners. International Journal of Applied Linguistics and English Literature, 5(6), 34–42.

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

Nation, P. (1990). Teaching and learning vocabulary. Newbury House.

O’Toole, C. (2022). Developing an assessment of vocabulary profi-
cency for Irish-English bilingual children: The Irish crosslin-
guistic lexical tasks. Journal of Multilingual and Multicultural Development. Advance online publication. https://doi.org/10.1080/01434632.2022.2036743

Paivio, A. (1986). Mental representations: A dual coding approach. Oxford University Press.

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learn-
ing through listening to songs. Studies in Second Language Acquisition, 44(4), 745–768.

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. TESOL Q., 53(4), 1008–1032.

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. Language Learning, 64(4), 878–912.

Polakova, P., & Klimova, B. (2022). Vocabulary mobile learning application in blended English language learning. Frontiers in Psychology, 13, 869055.

Ramezanali, N., & Faez, F. (2019). Vocabulary learning and retention through multimedia glossing. Language Learning & Technology, 23, 105–124.

Rassaei, E. (2018). Computer-mediated textual and audio glosses, perceptual style and L2 vocabulary learning. Language Teaching Research, 22(6), 657–675.

Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 2, 129–158.

Schmitt, N. (1997). Vocabulary learning strategies. In Schmitt, N., & McCarthy, M. (Eds.), Vocabulary: Description, acquisition, and pedagogy (pp. 199–227). Cambridge University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. Language Teaching Research, 12(3), 329–363.

Schnitz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representations. Learning and Instruction, 13, 141–156.

Schüller, A., & Mayer, M. G. (2020). Illustrations before text reduce visuospatial working memory load during text processing. Discourse Processes, 57(8), 627–658.

Sheshadri, S., Zhao, S. D., Chen, Y., & Fjeld, M. (2020, April 25–30). Learn with haptics: Improving vocabulary recall with free-form digital annotation on touchscreen mobiles [Conference session]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI ’20), Honolulu, HI, United States (pp. 1–13). ACM.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. In Schmitt, N., & McCarthy, M. (Eds.), Vocabulary: Description, acquisition, and pedagogy (pp. 156–180). Cambridge University Press.

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. Language Learning & Technology, 14(2), 50–73.

Tavakoli, M., & Gerami, E. (2013). The effect of keyword and pictorial methods on EFL learners’ vocabulary learning and retention. Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras, 19, 299–316.

Wang, S., & Lee, C. I. (2021). Multimedia gloss presentation: Learners’ preference and the effects on EFL vocabulary learning and reading comprehension. Frontiers in Psychology, 11, 3950.