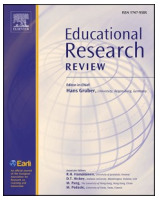
Educational Research Review 47 (2025) 100681

Contents lists available at ScienceDirect

Educational Research Review

journal homepage: www.elsevier.com/locate/edurev



Review

Effectiveness of L1 and pictures in multimedia conditions on learning second-language vocabulary: A meta-analysis

Caihui Zhang [a,b,d], Giovanni Sala [c,*], Fernand Gobet [d,e]

[a] Bilingual Cognition and Development Lab, Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, China
[b] Faculty of English Language and Culture, Guangdong University of Foreign Studies, China
[c] Department of Psychology, University of Liverpool, UK
[d] Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, UK
[e] School of Psychology, University of Roehampton, UK

ARTICLE INFO

Keywords:
Dual Coding Theory
Multimedia
Second-language acquisition
Vocabulary
Meta-analysis

ABSTRACT

Multimedia inputs have been often used in second language (L2) vocabulary learning; however, the effective elements in multimedia inputs for L2 vocabulary learning have hardly been established. This study aims to identify the effective element(s) and further clarifies the meaning of different "domains" in multimedia L2 vocabulary learning. Considering that the learning target (L2 vocabulary) belongs to the verbal domain, the meaningful inputs are then constructed as within-domain (i.e. L1-based), cross-domain (i.e. Picture-based), and mixed-domain (i.e. L1+picture-based) learning conditions. The present study firstly conducted an omnibus analysis and then three meta-analyses: (a) within-domain vs. cross-domain (20 studies, 51 effect sizes), (b) within-domain vs. mixed-domain (21 studies, 55 effect sizes), and (c) cross-domain vs. mixed-domain (9 studies, 27 effect sizes), for a total of 2056 participants. Both immediate and available delayed tests were used as dependent variables. Four moderators were used to identify potential predictors. The results indicate that, at the immediate tests, the mixed-domain condition consistently outperforms the within-domain condition ($g = 0.334$, $p < 0.001$) and cross-domain condition ($g = 0.350$, $p = 0.001$) in facilitating L2 vocabulary learning. However, at the delayed tests, the advantageous effect of the mixed-domain condition only marginally outperformed the cross-domain condition ($g = 0.271$, $p = 0.067$). No significant difference was found between the within-domain and cross-domain conditions in L2 vocabulary learning. No significant moderator was detected at either the immediate tests or the delayed tests. These findings highlight the benefit of incorporating both L1 words and pictures into L2 vocabulary learning.

1. Introduction

Learning a second language (L2) or a foreign language (FL) usually costs years of effortful hard work and yields different levels of proficiency. Given that vocabulary learning is typically the first step in L2 learning, a long-standing question in this field concerns the most effective way to learn L2 words.

According to Nation's (2001) framework, acquiring knowledge of an L2 word is a multifaceted process encompassing nine aspects related to its form, meaning, and use. When starting to learn a new word in a second language, the initial step is to learn its form (either

* Corresponding author.
E-mail address: Giovanni.Sala@liverpool.ac.uk (G. Sala).

https://doi.org/10.1016/j.edurev.2025.100681
Received 3 October 2023; Received in revised form 12 March 2025; Accepted 22 March 2025
Available online 22 March 2025

spoken or written form) and meaning, which is the foundation for learning the subsequent aspects of an L2 word. In L2 vocabulary learning, learning the forms of words is more challenging than learning their meanings, since there is more overlap in meaning between two distinct languages than in their shared word forms (Nation, 2001).

Compared with bilingual children, most adults might meet "fossilization" (Jiang, 2000) during L2 learning, which is a phenomenon where a learner's L2 stops developing despite continued exposure and practice. The Revised Hierarchical Model (RHM) proposed by Kroll and Stewart (1994) indirectly explained this phenomenon by the assumption that the conceptual link between L2 and concepts is weaker than that between L2 and first language (i.e., L1), which leads learners to rely too heavily on their L1 vocabulary and grammar, causing transfer errors that become ingrained in their L2 system. One possible reason leading to the strong link between L2 and L1 might be the ways of learning L2 vocabulary.

For most concrete L2 noun words, L2 learners can access their meanings through two main routes: verbally via L1 translation or non-verbally via images. Learning L2 words by L1 translation is typically regarded as a convenient learning method, especially for adults. Although the meanings of words in two languages may not be identical, most of the L2 word meanings can be found in their L1. With a dictionary on hand, L2 learners can access the meaning of a large number of L2 words rapidly.

Visual images, or pictures, are another way to convey meanings (Harrison, 2003) and are typical means for young children to learn a language. With the development of advanced technology, there has been a growing research interest in bringing multimedia into education, as well as into L2 vocabulary learning. According to Mayer and Fiorella (2022), "multimedia" can be defined as the presence of both verbal and pictorial materials. The verbal materials are presented in either spoken or written form, while the pictorial materials are presented in the form of illustrations, photos, animations, or videos.

The Cognitive Theory of Multimedia Learning proposed by Mayer and Fiorella (2022) builds upon Dual Coding Theory (Paivio, 1990), suggesting that information can be encoded via both visual and verbal channels. This simultaneous activation of verbal and pictorial representations fosters learning, as both types of information can serve as cues during retrieval. Consequently, they advocate "multimedia learning" as a means to construct mental representations from both words and pictures.

1.1. Mixed findings of multimedia inputs in L2 vocabulary learning

A large number of empirical studies have been conducted to verify the effectiveness of multimedia learning. Evidence supporting the Dual Coding Theory mainly comes from research in cognitive learning, which demonstrated that people learn complex concepts better when they are presented with both explanatory text and graphics than when they are presented with text alone (Mayer, 2002; Mayer & Moreno, 2003). However, the findings based on the predictions of the Dual Coding Theory in L2 vocabulary learning are rather mixed. Whilst some of the studies found the facilitation effect brought about by pictorial information (e.g., Boddaert et al., 2021; Emirmustafaoğlu & Gökmen, 2015; Liu et al., 2021; Morett, 2019), some studies supported that L1 translation would be more effective (e.g., Comesaña et al., 2012; Lotto & De Groot, 1998), and others failed to detect any difference between the two learning methods (e.g., Çakmak & Erçetin, 2018), despite the use of concrete L2 words in all cases. For example, Çakmak and Erçetin (2018) conducted a between-subject design study with 44 freshmen to learn English as an L2. The study compared four gloss input modes: textual-only, pictorial-only, and dual-channel (textual-plus-pictorial) gloss as well as a control condition where no glosses were provided. The results showed that the modes of gloss input did not affect learners' performance on L2 vocabulary recognition or production. Thus, they cast doubt on the hypothesis that increasing the number of multimedia elements improves L2 vocabulary learning.

To further examine the effect of multimedia input on L2 vocabulary learning, we distinguish the elements of multimedia input into two different domains: linguistic (e.g., L1 and L2) and non-linguistic (e.g., pictures). Based on the combination of inputs, domain conditions can be further classified as within-domain condition (i.e., learning L2 by L1), cross-domain condition (i.e., learning L2 by picture) and mixed-domain condition (i.e., learning L2 by L1 + picture). Although numerous studies found that the mixed-domain condition is more effective than the within-domain and cross-domain conditions (e.g., Warren et al., 2018; Çakmak & Erçetin, 2018), it is difficult to ascertain whether the facilitation effect is brought about by L1 translation, picture, or the combination of both. Therefore, it is worth separately investigating and discussing the effect induced by L1 translation (i.e., within-domain learning), picture (i.e., cross-domain learning), and L1 + picture (i.e., mixed-domain learning) in L2 vocabulary learning.

1.2. Potential moderators

Apart from the different domain conditions, some other potential factors might also influence the effectiveness of learning methods. Our analysis of the experimental design of previous studies found four potential factors that might affect learning: the learner, the learning target, the type of measurement, and the time of measurement.

First, with respect to the learner, age might play a key role in the success of L2 learning methods. For adults whose cognitive system and L1 literal system have been well-constructed, the findings for the effectiveness of L1-based and Picture-based learning methods were inconsistent. For instance, Bates and Son (2020) found that the Picture-based method was more effective whereas Alzahrani and Roberts (2021) found they were equally effective. Comparably, children cannot always rely on their L1 literal knowledge and may turn to visual information as effective meaning input. For example, Boddaert et al. (2021) compared L1-based and Picture-based learning methods in 3rd-grade children and found that the Picture-based L2 vocabulary learning method was more effective than the L1-based method. However, Comesaña et al. (2012) found an advantageous effect of the L1-based learning method in 7th-grade children. The inconsistent findings underscore the importance of classifying results for children and adults separately to accurately assess the effectiveness of L1-based and Picture-based learning methods. Therefore, when we talk about the effectiveness of a learning method, it would be worth investigating the potential effects brought about by the age of learners.

Second, concerning the learning target, the modality and form of target words might also affect learning performance. For most of the studies, the L2 written word was set as the learning target (e.g., Carpenter & Geller, 2020; Carpenter & Olson, 2012; Comesaña et al., 2009; Kost et al., 1999; Liu et al., 2021), with most of the remaining studies requiring participants to learn both L2 written word and audio (e.g., Alzahrani & Roberts, 2021; Bates & Son, 2020; Comesaña et al., 2012; Morett, 2019; Çakmak & Erçetin, 2018). We found only one study that took the L2 audio as the learning target (Boddaert et al., 2021). However, learning both the L2 written word and audio could overload learners' cognitive ability and might lead to an adverse effect. According to Sweller's (2005) Cognitive Load Theory, individuals' working memory is severely limited in dealing with novel information, and the input presented to learners should be carefully tailored, or it might lead to negative effects. Thus, more information is not necessarily better in L2 vocabulary learning. The effectiveness of learning methods should be discussed conditionally, depending on the content, form, and modality of the target learning language.

Third, the effectiveness of measurement could also influence the evaluation of the effectiveness of learning methods. Many different tasks have been adopted for assessing learners' word knowledge, such as word-meaning matching tasks (e.g., Boddaert et al., 2021), multiple-choice questions tasks (e.g., Liu et al., 2021), judgement tasks (e.g., Carpenter & Olson, 2012), recall tasks (e.g., Carpenter & Geller, 2020), translation tasks (e.g., Morett, 2019), and so on. Nation (2001) classified tasks measuring word knowledge into two types: receptive tasks and productive tasks. Receptive tasks require learners to recognize the correct relations between the target word and the options provided, whereas productive tasks require learners to produce word information according to the cue provided, either in spoken or written form. Most of the previous studies adopted either productive (e.g., Bates & Son, 2020; Carpenter & Geller, 2020; Morett, 2019) or receptive tasks (e.g., Boddaert et al., 2021; Lian et al., 2017; Liu et al., 2021), and few involved both productive and receptive tasks (e.g., Warren et al., 2018; Çakmak & Erçetin, 2018). Clearly, the type of task used to measure L2 vocabulary knowledge could influence the conclusions we draw. Therefore, we should consider the type of task used when evaluating the effectiveness of a learning method.

Last but not least, the timing of testing could also influence our evaluation of the effectiveness of learning methods. Compared to the tests conducted immediately after learning, delayed tests offer a better window to observe the effectiveness of learning conditions in the long term. Immediate tests might reflect short-term memory and initial learning, but delayed tests assess how well the information is consolidated and retained in long-term memory. For example, Comesaña et al. (2012) compared the L1-based learning condition and the Picture-based learning condition in facilitating children's L2 vocabulary learning, and found that children achieved better performance in the L1-based learning condition in the one-week-delayed tests, but the advantageous effect was not found in the immediate tests. Thus, the timing of delayed testing is another potential moderator that could influence our conclusion. Different delayed time lengths have been used in this study: one day (e.g., Shen, 2010), one week (e.g., Morett, 2019), two weeks (e.g., Lian et al., 2017), three weeks (e.g., Jones, 2004), and four weeks (e.g., Alzahrani & Roberts, 2021). Studies that conducted a delayed test make it possible to investigate the long-term effect of learning conditions. Importantly, the use of delayed testing aligns with real-world educational goals, where the aim is not just for students to learn information temporarily, but to retain and apply it long after the initial learning period.

1.3. Previous meta-studies

A few systematic reviews examined the effect of multimedia input on L2 vocabulary learning; however, they did not jointly discuss the effects occasioned by the within-domain (i.e., L1-based), the cross-domain (i.e., Picture-based), and the mixed-domain (i.e., L1 + picture-based) conditions. For example, Ramezanali et al. (2021), Vahedi et al. (2016), and Yun (2011) compared the within-domain and mixed-domain conditions. They found that the mixed-domain condition was more effective than the within-domain condition for L2 vocabulary learning. However, these studies took the number of input modes (i.e., single, dual, or triple) as the variable of interest rather than the types of input modes (i.e., textual or pictorial). Moreover, the visual input in their mixed-domain conditions varied from pictures to video clips. It is thus hard to tell whether the advantageous effect was brought about by the picture itself or the combination of L1, picture, video, and auditory sound. Huang (2012) and Yanagisawa et al. (2020) compared the effectiveness of within-domain and cross-domain conditions. However, due to the limited number of studies (only 8 studies) in Huang (2012) and the imbalanced number of effect sizes per condition in Yanagisawa et al. (2020)—3 effect sizes in the auditory condition, 8 in the pictorial condition, and 143 in the textual condition—they failed to detect any differences. Therefore, it is important to carry out an updated meta-analysis comparing the effectiveness of the within-domain, the cross-domain, and the mixed-domain conditions in L2 vocabulary learning.

In addition, some important moderators should be taken into consideration. Vahedi et al. (2016) conducted moderator analyses and found that the intensity of the program and the L2 proficiency level of learners are potential moderators influencing the heterogeneity between effect sizes. Moreover, Ramezanali et al. (2021) considered nearly 11 moderators in their meta-analysis, such as the quality of the data sample (e.g., journal or thesis), learners' characteristics (e.g., L2 proficiency, education), gloss features (e.g., the number of glosses, languages), text features (e.g., narrative, expository) and vocabulary measurement formats (e.g., form recall, form recognition, meaning recall, meaning recognition). The results showed that only the educational level of learners and the text features influenced the learners' performance. However, none of the previous meta-analyses mentioned above investigated how learners' age (i.e., children vs. adults), modality mode of the L2 (i.e., written form, spoken form, or both), type of testing tasks (i.e., receptive vs. productive), and timing of delayed testing (i.e., one day delayed, one week delayed, two weeks delayed, three weeks delayed, and four weeks delayed) influence heterogeneity. Thus, no definite conclusion can be drawn about the effectiveness of within-domain, cross-domain and mixed-domain conditions in facilitating L2 vocabulary learning, nor about the role of potential moderators.

Given the large amount of experimental evidence accumulated in the last few years, the theoretical and practical importance of

understanding how newly learned L2 words are mentally represented and accessed, and the contradictory conclusions reached by different researchers in the field, an up-to-date meta-analytic synthesis implementing a sound modelling design is required.

1.4. The present study

This paper evaluates the effectiveness of within-domain (i.e., L1-based), cross-domain (i.e., Picture-based) and mixed-domain (i.e., L1+picture-based) conditions on L2 vocabulary learning via meta-analysis. We focus on two primary goals. First, we test whether there are any differences across these three domain conditions (i.e., within-, cross-, and mixed-domain) at both immediate tests and delayed tests. We start by running an omnibus meta-analysis, that is, a meta-analysis including all the data. Then, we evaluate the effectiveness of different domain inputs on L2 vocabulary learning by conducting three meta-analyses (within-domain vs. cross-domain, within-domain vs. mixed-domain, and cross-domain vs. mixed-domain), with either immediate tests or delayed tests. These analyses will help us to understand the roles of L1 translation and picture in L2 vocabulary learning better, for both the short and long term.

Second, we aim to quantify and explain the amount of variability in the literature. We employ moderator analysis to investigate the potential sources of within- and between-study heterogeneity. This analysis addresses a fundamental point: statistically accounting for the degree of true heterogeneity is the only reliable way to make some sense of the mixed results the field has produced so far. Based on the literature reviewed in the introduction, we focus on the following moderators: participants' age (children vs. adults), modality mode of target L2 words (text vs. audio vs. text + audio), and type of testing task (receptive vs. productive). In the delayed tests, we added the timing of delayed testing (one day delayed vs. one week delayed vs. two weeks delayed vs. three weeks delayed vs. four weeks delayed) as another potential moderator. The moderator analysis will help us explain heterogeneity and examine the true effect induced by domain conditions.

2. Method

2.1. Literature search

A systematic search was employed to find the relevant studies (PRISMA statement; Moher et al., 2009). The following Boolean string was used: ("L2" OR "second language" OR "second-language" OR "foreign language" OR "foreign-language" OR "FL") AND ("word learning" OR "vocabulary learning" OR "vocabulary acquisition" OR "learning method*") AND ("experiment*" OR "participant*" OR "study") AND ("annotation*" OR "picture*" OR "pictorial" OR "image*" OR "photograph*" OR "text" OR "textual" OR "translation"). We searched through PsycINFO, Web of Science, Scopus, and ProQuest to identify all the potentially relevant studies. We retrieved 1299 records and removed 447 duplicates when records were synthesized in Endnote 20. Thus, 852 papers were left for further screening.

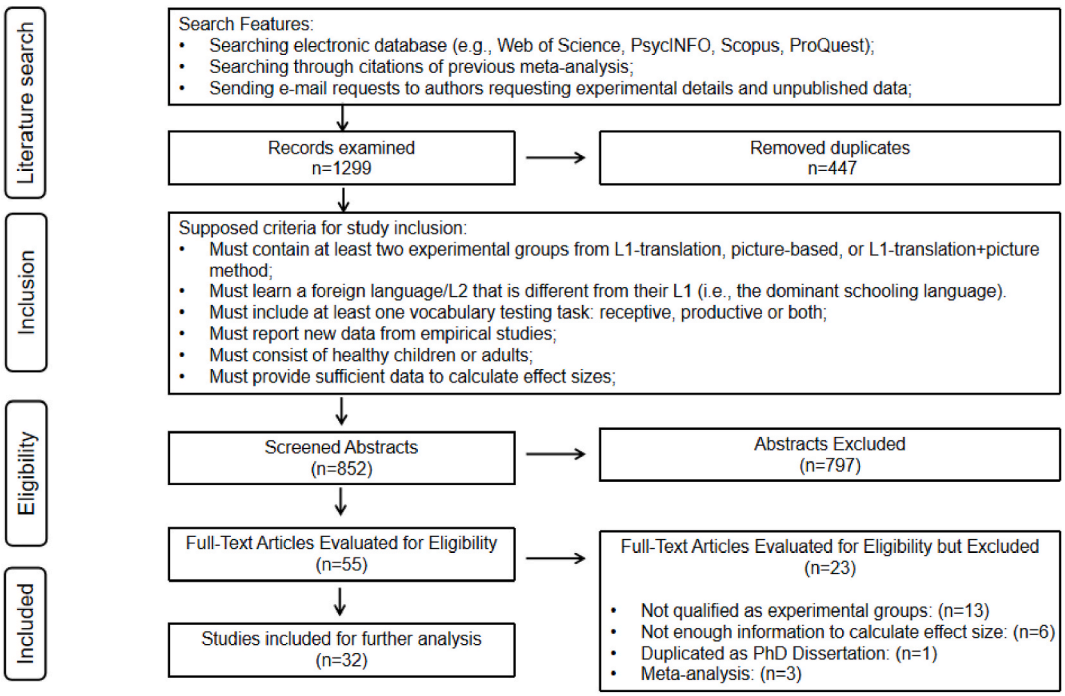


Fig. 1. Flow diagram of the search strategy.

2.2. Inclusion criteria

The studies were included according to the following six criteria:

1. The study conducted an experiment or quasi-experiment of foreign language vocabulary learning, containing at least two experimental groups that learnt either by within-domain condition (i.e., L1-based method), cross-domain condition (i.e., Picture-based method), or mixed-domain condition (i.e., L1+picture-based method). This criterion was fundamental to isolate the variable of interest, that is, the comparable impact of L1-based and Picture-based on the performance of second language vocabulary learning.
2. The participants must learn a foreign language/L2 which is different from their L1 (i.e., the dominant schooling language in their country).
3. At least one vocabulary testing task (either through receptive or productive assessment) was administered after the training. Tests could have been conducted immediately on the same day as the training, or delayed, taking place after the training day. Self-reported measures and parent/teacher rating questionnaires were excluded.
4. The study reported new data (i.e., it did not report duplicate results from previous studies).
5. The study focused on healthy learners, rather than people diagnosed with aphasia, hearing impairment, or learning disabilities.
6. The study contained appropriate quantitative and statistical data for calculating effect sizes and standard errors.

We searched for eligible published and unpublished papers through March 14th, 2022. We sent emails ($n = 28$) to researchers in the field asking for relevant information on experimental details (e.g., age, male rate, number of participants, procedure, and tasks testing vocabulary knowledge) and necessary data to calculate the effect sizes. We received eleven positive replies. In total, we found 32 studies conducted from 1992 to 2021 that met all the inclusion criteria. The entire procedure is described in Fig. 1. The supplemental materials available online contain the details of all the included studies and a list of the excluded studies (https://osf.io/he6pt/files/osfstorage).

2.3. Moderators

We assessed several moderators based on our review of the literature:

1. Age of learners: Quite a few studies did not report the mean age of their participants and roughly classified them as children and adults. Thus, we could not take age as a continuous variable in this study. According to Paulsen et al. (2011), starting from around 16 years of age, many adolescents exhibit a level of maturity and cognitive development that aligns more closely with adults than younger children; this can affect study outcomes in areas such as risk preference and cognitive development. Thus, we used 16 years as the cutoff for allocating participants from the previous studies into child and adult categories.
2. Modality and content of target learning words: Using the content and modality of target L2 words, we classified the learning target into L2 text, L2 audio, and Both (L2 text + L2 audio). However, only one study took L2 audio as a learning target. Thus, L2 audio was removed from the moderator analysis, leaving L2 text and Both as the categorical moderators.
3. Types of testing tasks: We classified all the testing tasks into receptive tasks and productive tasks according to the instructions. Specifically, receptive tasks mainly include word-meaning matching tasks, multiple-choice questions tasks, judgement tasks, and so on. By contrast, productive tasks refer to recall tasks, translation tasks, typing tasks, and so on.
4. Timing of delayed tests: We classified the timing of delayed tests into five groups: one day later, one week later, two weeks later, three weeks later, and four weeks later.

2.4. Effect size calculation

Effect size (ES) reflects the magnitude of the treatment effect (Borenstein et al., 2021). The ES (i.e., g in the formula below) was calculated based on the mean score (M) and the standard deviation (SD) of the eligible performance measure reported in the primary studies that met the inclusion criteria. The ES chosen for the analyses was Hedges's g, that is, the standardized mean difference corrected for the small-sample bias, as shown in the following formula.

$$g = d^* \left(1 - \frac{3}{4N - 9}\right)$$

with

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

where M_1 and M_2 refer to the mean performances of the two experimental groups; the groups consist of within-domain (i.e., L1-based L2 word learning), cross-domain (i.e., Picture-based L2 word learning), and mixed-domain (i.e., L1+picture-based L2 word learning) learning conditions in the three meta-analyses. SD_{pooled} is the pooled standard deviations of the two experimental groups, and N is the total sample size.

The sampling error variances were calculated with two formulas (Schmidt & Hunter, 2015, pp. 343–355) depending on the

experimental design implemented in the primary studies (i.e., between- or within-subject design). If the study followed a within-subject design, the sampling error variance was calculated by the following formula:

$$var_{g-within} = \frac{N-1}{N-3} * \frac{1}{N} * \left(1.33 + \frac{d^2}{2} * \frac{N}{N-1}\right) * \left(1 - \frac{3}{4N-9}\right)^2$$

If the study followed a between-subject design, the sampling error variance was calculated by the following formula:

$$var_{g\_between} = \frac{N-1}{N-3} * \frac{4}{N} * \left(1 + \frac{d^2}{8}\right) * \left(1 - \frac{3}{4N-9}\right)^2$$

2.5. Modeling approach

We classified all 32 studies into three meta-analysis datasets. Meta-analysis 1 compared the effectiveness of within-domain and cross-domain (i.e., L1-based vs. Picture-based) learning conditions. Meta-analysis 2 compared the effectiveness of within-domain and mixed-domain (i.e., L1-based vs. L1+picture-based) learning conditions. Meta-analysis 3 compared the effectiveness of cross-domain and mixed-domain (i.e., Picture-based vs. L1+picture-based) learning conditions.

We first ran an omnibus meta-analytic model—that is, a meta-analytic model including all the available effect sizes—for all 32 studies in Meta-analysis 1 (i.e., L1-based vs. Picture-based conditions) and Meta-analysis 2 (i.e., L1-based vs. L1+picture-based conditions). Meta-analysis 3 (i.e., Picture-based vs. L1+picture-based conditions) was not included because it was entirely composed of a subset of studies included in the other two meta-analyses. Then, a moderator analysis was run to test whether any moderator accounted for the difference among the three learning conditions.

The omnibus meta-analysis (Borenstein et al., 2021) was first conducted to provide a general overview of the effect size across all studies among L1-based, Picture-based and L1+picture-based learning conditions. After that, we conducted three meta-analyses separately to examine and compare the effectiveness of the L1-based, the Picture-based, and the L1+picture-based conditions pairwise. Outlier analysis and publication bias analysis were also conducted to guarantee the validity of the results.

These analyses were performed for both the immediate- and delayed-test effect sizes. The methods of running the omnibus analysis, the three pairwise meta-analyses, the outlier analysis, the moderator analysis, and the publication bias analysis are briefly explained in the following sections.

2.6. Analytic strategy

A systematic analytic strategy was then implemented for each meta-analysis (both the omnibus meta-analysis and the three meta-analyses grouped by condition). *Multilevel random-effect meta-analysis* was employed to run the intercept and meta-regression models (Viechtbauer, 2010). The intercept model was run first to estimate the overall effect sizes (ESs) and the true variances. The effect sizes extracted from one study were grouped into the same cluster. This technique allowed us to estimate both between-cluster and within-cluster true variance, that is, the amount of variance that was not due to sampling error (i.e., heterogeneity).

2.7. Outlier analysis

Sensitivity analyses were run in order to test the robustness of the results. Cook's distance was then calculated for each ES, and those ESs whose Cook's distance was greater than three times the mean were excluded. The intercept model was rerun without these ESs to test the robustness of the results and reduce spurious heterogeneity.

2.8. Moderator analysis

After removing the outliers, the moderator analysis was run. For the omnibus meta-analysis, we ran meta-regression analyses with four moderators (Condition, Age, Modality mode of L2 words, and Type of testing task). For the meta-analyses grouped by condition, the Condition moderator was dropped. Finally, for the meta-analyses with delayed tests, timing of delayed testing was added as a moderator.

In each set of analyses, all the possible meta-regression models with the above moderators were run and the model with the lowest Bayesian Information Criterion (BIC) was selected.

2.9. Publication bias analysis

Finally, funnel plots and the PET-PEESE method (Stanley, 2017) were employed to estimate the effect of publication bias on the results. Funnel plots depict the distribution of the effect sizes as a function of the squared root of their sampling error variance (i.e., their standard error). If publication bias is present, funnel plots should exhibit a marked asymmetry in the distribution of the effect sizes due to the systematic suppression of studies with low precision (i.e., small sample sizes) and effect sizes close to null. Otherwise, publication bias is probably absent. Likewise, PET-PEESE consists of two regression analyses relating effect sizes and standard errors (PET) and variances (PEESE). If standard errors or variances significantly predict the magnitude of effect sizes (i.e., the lower the

precision, the bigger the effect size), then publication bias is probably an issue.

Publication bias analysis was run only on the three meta-analyses grouped by condition. The rationale for this choice was that publication bias analysis usually does not perform accurately when considerable heterogeneity is present. In fact, we expected heterogeneity due to the differences in the distribution of the effect sizes depending upon the condition.

The metafor R package (Viechtbauer, 2010) was used to perform the analyses above. The R codes and results are available in the online supplementary material (https://osf.io/he6pt/files/osfstorage?view_only=2790863dd84f4d23ba15764b8577cf45). Due to space limitations, the description of the meta-analytical techniques was necessarily concise. For further details, see Borenstein et al. (2021).

3. Results

3.1. Descriptive statistics

For the immediate tests, there were 20 studies (51 effect sizes with 1164 participants) involved in Meta-analysis 1, which compared the L1-based and the Picture-based conditions in L2 vocabulary learning. In Meta-analysis 2, which compared the L1-based and the L1+picture-based conditions in L2 vocabulary learning, there were 21 studies (55 effect sizes with 1498 participants). In Meta-analysis 3, which compared the Picture-based and the L1+picture-based conditions in L2 vocabulary learning, there were 9 studies (27 effect sizes with 606 participants) involved.

However, not all studies included the delayed tests and the time of delay varied from one day later to one week later, two weeks later, three weeks later, and one month later. To be specific, 12 studies (25 effect sizes with 656 participants) conducted delayed tests in Meta-analysis 1, 17 studies (35 effect sizes with 957 participants) in Meta-analysis 2 (we excluded one effect size as the study did not include an immediate test), and 8 studies (16 effect sizes with 490 participants) in Meta-analysis 3.

3.2. Immediate-effect meta-analysis

3.2.1. Omnibus meta-analysis

The results of the omnibus meta-analysis for the immediate tests show that the overall effect size was significant ($g = 0.231$, $se = 0.057$, $p < 0.001$, $m = 32$, $k = 106$). This effect was still robust after removing the outliers ($g = 0.209$, $se = 0.048$, $p < 0.001$, $m = 30$, $k = 98$). The residual heterogeneity after accounting for the outliers and the effects of the moderators was low yet significant ($p < 0.001$). Between-study true variance was $\sigma^2 = 0.042$ while within-study true variance was $\sigma^2 = 0.000$.

All the possible combinations of moderators were evaluated in the moderator analysis. The best model according to the BIC showed that the only significant predictor was Condition. Meta-analysis 2 was significantly associated with higher effect sizes ($b = 0.287$, $se = 0.067$, $p < 0.001$).

3.2.2. The three meta-analyses grouped by condition

Consistent with the results of the moderator analysis in the omnibus meta-analysis, the meta-analyses grouped by condition showed different effect sizes.

The results of the three intercept models for the immediate tests show that the overall effect size of Meta-analysis 1 was not significant ($g = 0.122$, $se = 0.085$, $p = 0.171$, $m = 20$, $k = 51$), suggesting no significant differences between the L1-based and the Picture-based learning conditions in the immediate tests. A robust overall effect size was found in Meta-analysis 2 ($g = 0.343$, $se = 0.066$, $p < 0.001$, $m = 21$, $k = 55$), indicating that the L1+picture-based condition outperformed the L1-based condition in the immediate tests. Lastly, a significant overall effect size was found in Meta-analysis 3 ($g = 0.393$, $se = 0.093$, $p = 0.003$, $m = 9$, $k = 27$), indicating that the L1+picture-based condition also outperformed the Picture-based condition in the immediate tests.

Table 1
The results of the three intercept models of immediate tests after removing outliers.

Model	Meta-analysis 1: L1 vs. Picture	Meta-analysis 2: L1 vs. L1+Picture	Meta-analysis 3: Picture vs. L1+Picture
m	18	20	8
k	47	51	24
g	0.072	0.334	0.350
se	0.055	0.066	0.066
t	1.319	5.052	5.268
p	0.205	<0.001***	0.001**
95 % lower bound *CI*	-0.043	0.196	0.193
95 % upper bound *CI*	0.188	0.472	0.507
Between-study true variance	0.016	0.058	0.000
Within-study true variance	0.014	0.000	0.020
Q.p	0.027*	<0.001***	0.215

Note. m = numbers of studies; k = number of effect sizes; g = overall effect size; se = overall effect size's standard error; t = t-value statistics; p = overall effect size's significance; *CI*: confidence interval; *Q.p* = heterogeneity significance. ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$.

3.2.3. Outlier analysis

These results were robust to the effect of the outliers. The overall effect size of Meta-analysis 1 decreased from 0.122 ($se = 0.085$, $m = 20$, $k = 51$) to 0.072 ($se = 0.055$, $m = 18$, $k = 47$) and was still not significant ($p = 0.205$), suggesting no significant differences between the L1-based and the Picture-based learning conditions in the immediate tests. The overall effect size of Meta-analyses 2 ($g = 0.334$, $se = 0.066$, $p < 0.001$, $m = 20$, $k = 51$) and Meta-analysis 3 ($g = 0.350$, $se = 0.066$, $p = 0.001$, $m = 8$, $k = 24$) after removing outliers did not change much compared to the main analysis, indicating that the L1+picture-based condition outperformed the L1-based condition and the Picture-based condition in the immediate tests (Table 1).

3.2.4. Moderator analysis

As we found that the tests for heterogeneity were significant for Meta-analysis 1 and 2 in the immediate tests, we further explored the potential moderators for the two meta-analyses. The moderator analysis took Age, Modality mode of L2 words, and Type of testing task into the multivariate meta-analysis model. However, only the intercept model was selected and no other moderators predicted any effect for the two Meta-analyses. That is, no moderator exerted any noticeable effect in the immediate tests.

3.2.5. Publication bias analysis

Finally, all the publication bias analyses estimated a substantially null effect of publication bias. Neither the inspection of the funnel plots nor the PET-PEESE results showed any sign of publication bias. The funnel plots (Fig. 2) showed no sign of asymmetry in the distribution of the effect sizes. Consistent with the funnel plots, neither the standard errors (PET) nor the sampling error variances (PEESE) were significantly related to the ESs in any of the three meta-analyses (all $ps \geq 0.212$). Thus, we conclude that publication bias was not a concern in any of the meta-analyses.

3.3. Delayed-effect meta-analysis

3.3.1. Omnibus meta-analysis

The results of the omnibus meta-analysis for the delayed tests show that the overall effect size was not significant ($g = 0.164$, $se = 0.083$, $p = 0.065$, $m = 19$, $k = 59$). Removing outliers resulted in a slightly significant effect ($g = 0.201$, $se = 0.082$, $p = 0.025$, $m = 18$, $k = 56$). The residual heterogeneity was more pronounced than for the immediate-test models. Between-study true variance and within-study true variance were $\sigma^2 = 0.091$ and $\sigma^2 = 0.006$, respectively.

All the possible combinations of moderators were evaluated in the moderator analysis. The best model according to the BIC showed that there were no significant predictors (i.e., the intercept model was deemed the best).

3.3.2. The three meta-analyses grouped by condition

The results of the three intercept models for the delayed tests show that the overall effect size of Meta-analysis 1 was not significant ($g = 0.080$, $se = 0.117$, $p = 0.510$, $m = 11$, $k = 25$), suggesting no significant differences between the L1-based and the Picture-based learning conditions in the delayed tests. A significant overall effect size was found in Meta-analysis 2 ($g = 0.283$, $se = 0.090$, $p = 0.007$, $m = 15$, $k = 34$), indicating that the L1+picture-based condition outperformed the L1-based condition in the delayed tests. Lastly, a positive yet non-significant overall effect size was found in Meta-analysis 3 ($g = 0.187$, $se = 0.133$, $p = 0.210$, $m = 7$, $k = 16$), indicating

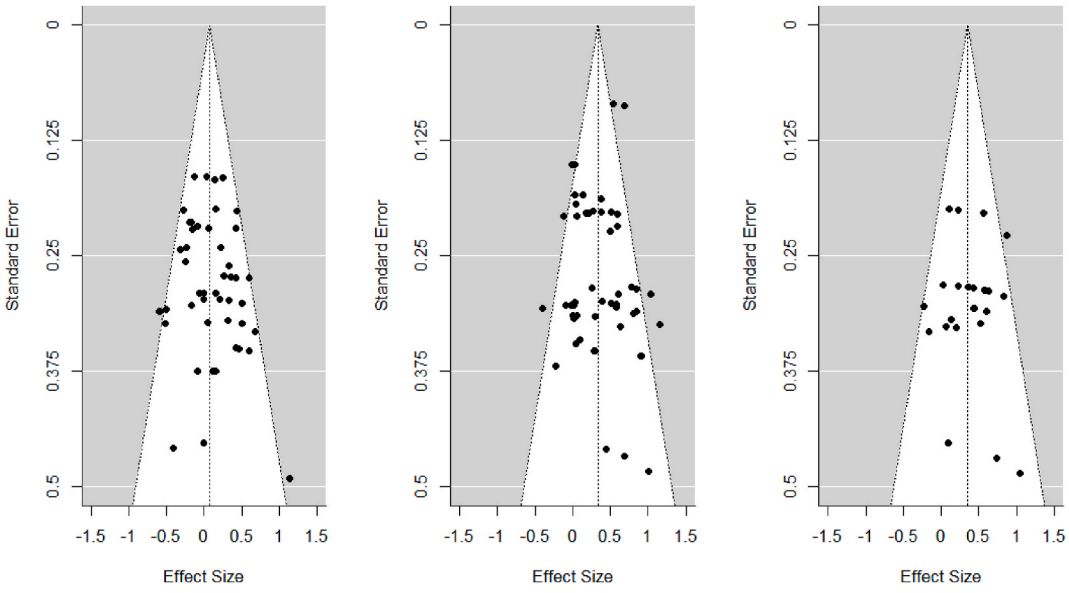


Fig. 2. The funnel plots of the three meta-analyses at the immediate tests.

Table 2
The results of the three intercept models of delayed tests after removing outliers.

Model	Meta-analysis 1: L1 vs. Picture	Meta-analysis 2: L1 vs. L1+Picture	Meta-analysis 3: Picture vs. L1+Picture
m	10	15	6
k	23	33	15
g	0.055	0.187	0.271
se	0.089	0.161	0.116
t	0.609	1.164	2.334
p	0.558	0.264	0.067
95 % lower bound *CI*	-0.148	-0.158	-0.028
95 % upper bound *CI*	0.257	0.533	0.569
Between-study true variance	0.047	0.338	0.048
Within-study true variance	0.000	0.005	0.000
Q.p	0.026*	<0.001***	0.149

Note. m = numbers of studies; k = number of effect sizes; g = overall effect size; se = overall effect size's standard error; t = t-value statistics; p = overall effect size's significance; *CI*: confidence interval; *Q.p* = heterogeneity significance. ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$.

that the L1+picture-based condition did not outperform the Picture-based condition in the delayed tests.

3.3.3. Outlier analysis

These results were a little bit different after removing the outliers. The overall effect size of Meta-analysis 1 was not significant ($g = 0.055$, $se = 0.089$, $p = 0.558$, $m = 10$, $k = 23$). The overall effect size of Meta-analyses 2 declined to non-significant ($g = 0.187$, $se = 0.161$, $p = 0.264$, $m = 15$, $k = 33$). Finally, the overall effect size of Meta-analysis 3 approached significance ($g = 0.271$, $se = 0.116$, $p = 0.067$, $m = 6$, $k = 15$). Thus, there was no significant difference between L1+picture-based condition, Picture-based, and L1-based learning conditions in the delayed tests (Table 2).

3.3.4. Moderator analysis

As we detected that the tests for heterogeneity were significant for Meta-analysis 1 and 2 in the delayed tests, we further explored the potential moderators for these two meta-analyses. The moderator analysis selected only one predictor (Age) in Meta-Analysis 2 but it did not reach significance ($b = 0.644$, $p = 0.152$). No moderator was detected in Meta-Analysis 1. The results showed that Age might be a potential moderator to influence the learning effectiveness between L1+picture-based and L1-based learning conditions in the delayed tests.

3.3.5. Publication bias analysis

Finally, all the publication bias analyses estimated a substantially null effect of publication bias. Just like in the immediate-test

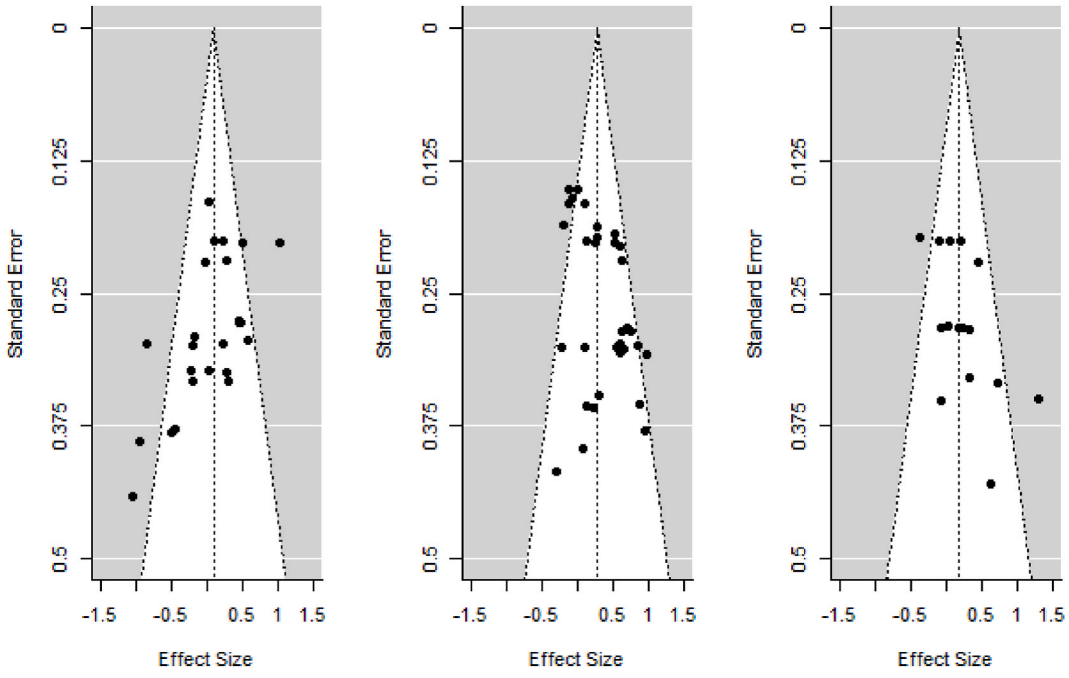


Fig. 3. The funnel plots of the three meta-analyses at the delayed tests.

models, neither the inspection of the funnel plots nor the PET-PEESE results showed any sign of publication bias. The funnel plots (Fig. 3) showed no apparent asymmetry in the distribution of the effect sizes. Consistent with the funnel plots, neither the standard errors (PET) nor the sampling error variances (PEESE) were significantly related to the ESs in any of the three meta-analyses (all $ps \geq 0.162$).

4. Discussion

The present study aimed to investigate the effectiveness of domain conditions on L2 vocabulary learning by conducting meta-analyses of previous studies. The research aimed to identify (a) the most effective semantic input for L2 vocabulary learning from within-domain (i.e., L1-based), cross-domain (i.e., Picture-based), or mixed-domain (i.e., L1+picture-based) learning conditions; as well as (b) the potential moderators that might influence the multimedia input of L2 vocabulary learning outcomes. To address the two research questions, this study compared the effectiveness of L1 translation, pictures, and the combination of L1 translation + picture at both immediate tests and delayed tests. An omnibus analysis was first conducted to provide a general overview of the effect size across all studies among L1-based, Picture-based and L1+picture-based learning conditions; then, three meta-analyses grouped by conditions were carried out to specify the differences: (a) within-domain vs. cross-domain; (b) within-domain vs. mixed-domain; and (c) cross-domain vs. mixed-domain. Moderator analyses were also conducted to examine how learners' age, modality modes of target L2 words, types of testing tasks, and timing of delayed testing influence the heterogeneity of the learning results. The results indicated that, on immediate tests, the mixed-domain condition outperformed both the within-domain and cross-domain conditions in facilitating L2 vocabulary learning. On the delayed tests, the mixed-domain condition marginally outperformed the cross-domain condition. No moderator effect was found in either the immediate tests or the delayed tests.

4.1. Mixed-domain learning condition facilitates L2 vocabulary learning

On the immediate tests, the findings of meta-analyses 2 and 3 showed that the mixed-domain condition outperformed the within-domain and the cross-domain learning conditions. On the delayed tests, the mixed-domain condition marginally outperformed the cross-domain condition. The results indicate that the L1+picture-based learning condition was more effective in facilitating L2 vocabulary learning than the L1-based and Picture-based learning conditions, although the effect was less clear with the delayed tests.

These findings are consistent with the predictions of the Dual Coding Theory (Paivio, 1990) and the Cognitive Theory of Multimedia Learning (Mayer & Fiorella, 2022), which suggest that learning materials combining both text and images will lead to better comprehension and retention than materials presented with text alone. In this study, we also found that the mixed-domain condition (i.e., L1+picture-based) outperformed not only the within-domain condition (i.e., L1-based), but also the cross-domain condition (i.e., Picture-based). As outlined in these theories, information from both verbal and pictorial channels could be processed simultaneously and serve as retrieving cues during vocabulary testing, thus fostering learning (Paivio & Csapo, 1973). Although the Cognitive Theory of Multimedia Learning (Mayer, 2022) mainly applies to complex material (e.g., learning about the brake systems), it can also be used for simpler tasks like L2 vocabulary learning, which requires less cognitive load. Arguably, the Picture-based learning condition also met the requirements of the suggested combination by providing both the picture and the word (L2 words) during the learning phase. However, the provided L2 words are novel information for learners, which cannot be processed as meaningful cues at the beginning of the learning phase. Therefore, learners in the Picture-based learning condition received only one memory cue, which was less effective than the two memory cues provided in the L1+picture-based learning condition. Similarly, the L1-based learning condition provided only one cue from the verbal memory, which was less effective than the L1+picture-based learning condition in facilitating L2 vocabulary learning.

An additional explanation for the weaker efficiency of pictures alone is that they can lead to ambiguity in word naming, or multiple possible interpretations (Vitkovitch & Tyrrell, 1995). The objects in the picture can not only refer to nouns but also adjectives or even verbs. For example, a picture of a well-dressed little girl can not only lead to the word "girl", but also "beautiful", "cute", or even "standing". Such ambiguity highlights how visual images might trigger different meanings depending on the context and the viewer's perceptual processes. This is particularly relevant in areas like visual learning, where the clarity of images becomes crucial to avoid misunderstandings or misinterpretations. Thus, L1 translations in the mixed-domain condition would be necessary to clarify the meaning of the pictures, and to help learners understand the meanings of the target L2 words.

The findings suggest that the most efficient way of learning L2 vocabulary is by both L1 translations and pictures, even for the group of adults. We understand that most adults prefer to learn L2 vocabulary based on L1 translations for the sake of convenience. However, this method would strengthen the link between L2 words and their L1 translations, which is weak compared with the link between L2 words and concepts (pictures/images can help visualize concepts). Using this weak link might lead to "fossilization" (Jiang, 2000), stopping learners from reaching high L2 proficiency levels (Kroll & Stewart, 1994). Thus, we strongly suggest that learners should build direct links between L2 vocabulary and concepts, a process that can be facilitated for adults by using both L1 translations and pictures.

4.2. No difference between within- and cross-domain conditions

One of the main goals of this study was to evaluate the effectiveness of within- and cross-domain conditions in L2 vocabulary learning, which was examined by comparing effectiveness between the L1-based and Picture-based conditions. However, the effect size in this meta-analysis was not statistically significant, either in the immediate tests or delayed tests, indicating that there was no

noteworthy difference between the L1-based and Picture-based conditions in facilitating L2 vocabulary learning.

A main limitation of this meta-analysis concerns the relatively small number of studies found in this field, as only 20 studies examined the effectiveness of within- and cross-domain conditions in L2 vocabulary learning. Most of the studies we collected in this meta-analysis (i.e., Meta-analysis 1) took L2 written words as the learning targets, and only one study (Boddaert et al., 2021) took L2 spoken words as the learning targets. Thus, we preferred to draw our conclusion conservatively on L2 written word learning and call for future studies to investigate L2 spoken word learning to make our understanding more comprehensive.

Moreover, nearly half of these studies were set to compare three learning conditions: L1-based, Picture-based, and L1+picture-based. As most of the studies highlight the facilitation effect brought by the L1+picture-based condition, the studies comparing the difference between the L1-based and the picture-based learning condition were too few to allow any robust conclusion. As reported by Carpenter and Olson (2012), learners' belief and confidence in a learning method could affect their learning outcomes. Undoubtedly, the L1+picture-based condition provides a better learning experience (e.g., Çakmak & Erçetin, 2018; Lian et al., 2017), and learners' common sense would also show more confidence in this learning condition, particularly for those experiments adopting within-subject design. The actual differences between the L1-based and the Picture-based learning conditions require additional investigations. For example, future studies could use neuroscience techniques to investigate the brain regions and connectivity underlying the different learning conditions. Behavioral performance merely reflects the outcomes of learning, while neural evidence offers another perspective for observing the neural foundations that underpin this performance.

4.3. Moderators in domain learning conditions

To our disappointment, we did not find any significant moderators in different domain learning conditions, neither in the immediate tests nor the delayed tests. Several reasons might account for this result.

First, the age moderator in this study was categorical rather than numerical. This decision stemmed from the fact that, in the data we collected, most studies only reported the age range of their participants, rather than providing detailed statistics like the specific mean age and standard deviations. Such data limitations make it difficult to treat age as a continuous moderator. Thus, we had to roughly divide our participants into "Children" and "Adults" in our meta-analyses. This rough classification might have stopped us from detecting any reliable moderating effect.

Second, the sample size for different modality modes of target L2 words was imbalanced. As noted above, we found only one study (Boddaert et al., 2021) that adopted L2 audio words as the learning target, and this category was removed in the moderator analysis. Thus, due to limited samples, our conclusions must be prudent concerning the modality mode of target L2 words.

As for the type of testing tasks, our results showed that participants' learning performance was not influenced by either receptive or productive tasks. One possible reason is that productive and receptive tasks are not mutually exclusive but often involve overlapping cognitive processes, and both require lexical access and mental representations of vocabulary. Our findings were similar to Çakmak and Erçetin's (2018) in that there was no significant difference in learners' performance on L2 vocabulary recognition and production. Nation and Meara (2013) also argued that learners' receptive knowledge of words can aid productive use, as they both involve recalling word meanings, forms and uses. Another potential reason is that the test formats of receptive tasks varied between studies, which might have affected our results. Generally, we classified all the testing tasks into receptive tasks or productive tasks. However, there were many different test formats for testing learners' receptive word knowledge. For example, in the study of Liu et al. (2021), the pictures used in the testing phase were the same as those in the learning phase. However, Boddaert et al. (2021) used different pictures in the testing phase referring to the same learning objects, which aimed to test learners' transformed knowledge. The variation of testing formats in receptive tasks might have influenced the results of identifying testing task types as a potential moderator. Future studies in this field could further explore the potential influence exerted by test formats.

Lastly, we did not detect any moderator effect from the timing of the delayed tests. An interesting finding was that the residual heterogeneity of the delayed-test model was more pronounced than the immediate-test model, which might be due to inconsistencies in selecting the timings of the delayed tests. Owing to the limited number of studies examining the delayed testing effect, the meta-analysis may lack sufficient power to detect a statistically significant moderator effect.

4.4. Limitations and future studies

Conducting the meta-analyses required coding the research designs and experimental paradigms of previous studies. This allowed us to identify several limitations that provide inspiration for future studies in this field.

Firstly, there is a dearth of studies investigating L2 spoken form learning as target learning modality. As noted above, we found only one study that examined the effectiveness of domain conditions in L2 spoken form learning (Boddaert et al., 2021). Due to the lack of research in this area, we were unable to include the "audio" category of modality in the moderator analysis. Relatively, it might be more difficult to investigate L2 spoken form learning than L2 written form learning, due to technical and platform constraints. However, L2 spoken form learning is a crucial part of L2 vocabulary learning that should not be ignored. Future studies should examine L2 spoken form learning to provide a more comprehensive understanding of multimedia L2 vocabulary learning.

Secondly, the evidence supporting the effectiveness of Picture-based conditions for L2 vocabulary learning in adults is mixed when compared with the facilitation effect observed in children (e.g., Fidler, 2011; Lian et al., 2017). As globalization continues to progress, an increasing number of adults seek to learn a second or third language to broaden their horizons. The potential market for adults' L2 learning is significant. Adults, with higher cognitive abilities and motivation than children, can assess word meanings using both L1 glosses and pictures. Future studies should investigate how adults benefit from the two domains in L2 vocabulary learning, providing

more practical implications for education in adults' L2 learning.

Thirdly, most of the results obtained in the meta-analyses were based on learners' retrieval performance, with fewer studies investigating the encoding process and how the domain conditions influence it. Understanding the encoding process is crucial for gaining a complete picture of L2 vocabulary learning, which may provide valuable insights into how different modality inputs affect the acquisition and retention of new vocabulary. By examining the encoding process of L2 vocabulary learning, researchers may be able to develop more effective teaching methods that capitalize on the strengths of different input modalities, especially in the context of language education where time and resources are often limited.

Fourthly, the distance between learners' L1 language and target language could be further considered. For example, although both Persian and Arabic use scripts derived from the Arabic script, Persian belongs to the Indo-Iranian branch of the Indo-European language family, whereas Arabic belongs to the Semitic branch of the Afro-Asiatic language family. Thus, for English native speakers, learning the sounds of Persian is typically less challenging than learning the Persian script. The more similar the learners' L1 and target languages are, the easier the learning process will be. Future studies should consider whether learners' L1 and target language come from the same language system in both written forms and spoken forms, which might also influence the evaluation of learning methods.

Last but not least, the current studies which investigated the delayed effect of a learning method are still few, which restricted our exploration on the delayed effect of learning methods in this study. Due to this limited number, our findings in the delayed tests were marginally significant. Most studies only conducted an immediate posttest after the learning phase, with no follow-up study to determine if the learned L2 word knowledge had been consolidated into long-term memory. The ultimate goal of L2 learning is not to store information in temporary memory but to store it in long-term memory. Therefore, future experimental studies that examine the effectiveness of a learning method should include some delayed posttests, such as one week or one month after the learning phase, which would help to understand the effectiveness of domain conditions in the long term.

5. Conclusion

In this study, we examined the effectiveness of within-domain condition (i.e., L1-based method), cross-domain condition (i.e., Picture-based method), and mixed-domain condition (i.e., L1+picture-based method) in multimedia L2 vocabulary learning with respect to both immediate posttest and delayed posttest. The results significantly corroborate and extend the conclusion reached in previous meta-analyses (e.g., Vahedi et al., 2016; Yun, 2011). First, the mixed-domain condition (i.e., L1+picture-based learning) outperformed the within-domain (i.e., L1-based learning) and the cross-domain (i.e., Picture-based learning) in facilitating L2 vocabulary learning, especially in the immediate tests of L2 written form learning. By contrast, no difference was found between the L1-based and Picture-based methods in L2 vocabulary learning. No significant moderator was found in this study, but the limitations identified from previous studies should be addressed in future studies.

CRediT authorship contribution statement

Caihui Zhang: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Giovanni Sala:** Methodology, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Fernand Gobet:** Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT 3.5 in order to check for typos and grammatical mistakes. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declarations of interest

None.

Acknowledgements

We thank the researchers who send us the requested information about the original studies. Caihui Zhang was supported by a fellowship from the Guangzhou Elites Scholarship Council and research funds from Center for Linguistics and Applied Linguistics at Guangdong University of Foreign Studies (NO: 201-X5224220).

Data availability

The data and R code are available at the following link: https://osf.io/he6pt/files/osfstorage.

References

Alzahrani, S., & Roberts, L. (2021). The effect of visuospatial designing elements of zoomable user interfaces on second language vocabulary acquisition. *System, 96*, Article 102396. https://doi.org/10.1016/j.system.2020.102396

Bates, J., & Son, J. B. (2020). English vocabulary learning with simplified pictures. *Teaching English as a Second or Foreign Language - Electronic Journal, 24*(3), 1–20.

Boddaert, G., Casalis, S., & Mahé, G. (2021). Photograph method fosters direct access to second-language word meaning: Direct evidence from a word-picture matching task. *British Journal of Developmental Psychology, 39*(3), 407–423. https://doi.org/10.1111/bjdp.12375

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Çakmak, F., & Erçetin, G. (2018). Effects of gloss type on text recall and incidental vocabulary learning in mobile-assisted L2 listening. *ReCALL, 30*(1), 24–47. https://doi.org/10.1017/S0958344017000155

Carpenter, S. K., & Geller, J. (2020). Is a picture really worth a thousand words? Evaluating contributions of fluency and analytic processing in metacognitive judgements for pictures in foreign language vocabulary learning. *Quarterly Journal of Experimental Psychology, 73*(2), 211–224. https://doi.org/10.1177/1747021819879416

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(1), 92. https://doi.org/10.1037/a0024828

Comesaña, M., Perea, M., Piñeiro, A., & Fraga, I. (2009). Vocabulary teaching strategies and conceptual representations of words in L2 in children: Evidence with novice learners. *Journal of Experimental Child Psychology, 104*(1), 22–33. https://doi.org/10.1016/j.jecp.2008.10.004

Comesaña, M., Soares, A. P., Sanchez-Casas, R., & Lima, C. (2012). Lexical and semantic representations in the acquisition of L2 cognate and non-cognate words: Evidence from two learning methods in children. *British Journal of Psychology, 103*(3), 378–392. https://doi.org/10.1111/j.2044-8295.2011.02080.x

Emirmustafaoğlu, 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. https://doi.org/10.1016/j.sbspro.2015.07.559

Fidler, J. (2011). *Evaluating the effectiveness of multimedia annotation modes and vocabulary learning tasks for second Language Vocabulary acquisition. Master thesis.* University of Haifa (Israel).

Harrison, C. (2003). Visual social semiotics: Understanding how still images make meaning. *Technical Communication, 50*(1), 46–60.

Huang, S. F. (2012). *Effects of tasks and glosses on L2 incidental vocabulary learning: Meta-analyses*. Doctoral dissertation, Texas A & M University.

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics, 21*(1), 47–77. https://doi.org/10.1093/applin/21.1.47

Jones, L. (2004). Testing L2 vocabulary recognition and recall using pictorial and written test items. *Language, Learning and Technology, 8*(3), 122–143.

Kost, C. R., Foss, P., & Lenzini Jr, J. J. (1999). Textual and pictorial glosses: Effectiveness on incidental vocabulary growth when reading in a foreign language. *Foreign Language Annals, 32*(1), 89–97. https://doi.org/10.1111/j.1944-9720.1999.tb02378.x

Kroll, J. F., & Stewart, E. (1994). Category Interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Lian, Y. Y., Chen, C. M., & Li, Y. R. (2017). Effects of collaborative multimedia annotations on elementary school students' vocabulary learning performance. In *2017 6th IIAI international congress on advanced applied informatics (IIAI-AAI)* (pp. 649–654). IEEE. https://doi.org/10.1109/IIAI-AAI.2017.52.

Liu, X. Y., Horinouchi, H., Yang, Y. T., Yan, Y., Ando, M., Obinna, U. J., Namba, S. S., & Kambara, T. (2021). Pictorial referents facilitate recognition and retrieval speeds of associations between novel words in a second language (L2) and referents. *Frontiers in Communication, 6*. https://doi.org/10.3389/fcomm.2021.605009. Article 605009.

Lotto, L., & De Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning, 48*(1), 31–69. https://doi.org/10.1111/1467-9922.00032

Mayer, R. E. (2002). Multimedia learning. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 41, pp. 85–139). Academic Press. https://doi.org/10.1016/S0079-7421(02)80005-6.

Mayer, R. E. (2022). Cognitive theory of multimedia learning. In R. E. Mayer, & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 57–72). Cambridge University Press.

Mayer, R. E., & Fiorella, L. (2022). *The Cambridge handbook of multimedia learning* (3rd ed.). Cambridge University Press.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Morett, L. M. (2019). The power of an image: Images, not glosses, enhance learning of concrete L2 words in beginning learners. *Journal of Psycholinguistic Research, 48*(3), 643–664. https://doi.org/10.1007/s10936-018-9623-2

Nation, I. S. P. (2001). *Learning vocabulary in another language* (Vol. 10). Cambridge University Press.

Nation, P., & Meara, P. (2013). 3 vocabulary. In *An introduction to applied linguistics* (pp. 44–62). Routledge.

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.

Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology, 5*(2), 176–206. https://doi.org/10.1016/0010-0285(73)90032-7

Paulsen, D. J., Platt, M. L., Huettel, S. A., & Brannon, E. M. (2011). Decision-making under risk in children, adolescents, and young adults. *Frontiers in Psychology, 2*, 72. https://doi.org/10.3389/fpsyg.2011.00072

Ramezanali, N., Uchihara, T., & Faez, F. (2021). Efficacy of multimodal glossing on second language vocabulary learning: A meta-analysis. *Tesol Quarterly, 55*(1), 105–133. https://doi.org/10.1002/tesq.579

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Shen, H. H. (2010). Imagery and verbal coding approaches in Chinese vocabulary instruction. *Language Teaching Research, 14*(4), 485–499. https://doi.org/10.1177/1362168810375370

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8*, 581–591. https://doi.org/10.1177/1948550617693062

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19–30). https://doi.org/10.1017/cbo9780511816819.003

Vahedi, V. S., Ghonsooly, B., & Pishghadam, R. (2016). Vocabulary glossing: A meta-analysis of the relative effectiveness of different gloss types on L2 vocabulary acquisition. *Teaching English with Technology, 16*(2), 3–25.

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software, 36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Vitkovitch, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *The Quarterly Journal of Experimental Psychology, 48*(4), 822–848. https://doi.org/10.1080/14640749508401419

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

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading?: A meta-regression analysis. *Studies in Second Language Acquisition, 42*(2), 411–438. https://doi.org/10.1017/S0272263119000688

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning, 24*(1), 39–58. https://doi.org/10.1080/09588221.2010.523285