



The effects of multimedia gloss on incidental vocabulary learning: A meta-analysis

 QIWEI YANG^a  

School of Foreign Languages and Literature, Beijing Normal University

Abstract

Technological advancements have heightened interest in multimedia glosses, including pictorial and audio glosses. However, debates persist regarding their impact on incidental vocabulary learning. While many studies highlight the positive effects of multimedia glosses, others offer opposing views. To address these conflicting findings, a comprehensive approach like meta-analysis is necessary, as it can synthesize varying results from multiple studies to provide a more definitive conclusion. The present study employs meta-analysis to assess the overall impact of multimedia glosses on incidental vocabulary learning and identify factors that might moderate these effects. Analyzing 22 studies with 84 effect sizes, the results reveal an overall effect size of 0.727 from a random effects analysis. When combining immediate and delayed tests, this effect size increases to 0.882. The heterogeneity test indicates that studies using L2 glosses, receptive tests, distractors, and nonwords tend to yield larger effect sizes. Furthermore, meta-regression results show that only the nonword condition and glosses language significantly influences the model. The study discusses implications and limitations, providing insights for future research.

Keywords: incidental vocabulary learning; multimedia glosses; meta-analysis

Introduction

The incremental nature of vocabulary learning requires sustained effort, as building a robust vocabulary is essential for language proficiency. Incidental vocabulary learning, which occurs through exposure to language input, complements intentional vocabulary learning by gradually enhancing vocabulary knowledge (Nation, 2001; Schmitt, 2010). Despite its potential benefits, incidental vocabulary learning has been criticized for its inefficiency in a foreign language when learners rely solely on reading a single book or engaging in short-term exposure (Teng & Reynolds, 2024).

To improve the efficiency of incidental vocabulary learning, glosses—annotations or explanations accompanying text—are often used. These glosses assist in comprehending the text and facilitate vocabulary acquisition (Mohsen & Balakumar, 2011). Research has explored various aspects of glosses, including their effectiveness in enhancing reading comprehension and vocabulary acquisition. For example, Taylor (2013) demonstrated that glosses could significantly improve reading comprehension, while Hulstijn et al. (1997) focused on their role in vocabulary acquisition. The language used in glosses, whether the learner's native language or the target language, has also been extensively studied. Yoshii (2006) found that the choice of language in glosses can impact learning outcomes, with some studies suggesting that glosses in the target language may lead to better vocabulary retention. Furthermore, the mode of glosses, such as pictorial or audio, has been a topic of interest. Rassaei (2018) investigated the effectiveness of pictorial glosses, finding that they can enhance vocabulary learning by providing visual context. However, the integration of multimedia elements like pictures in glosses has led to mixed results. For instance, Boers et al. (2017) reported that adding pictures might sometimes impede vocabulary acquisition, possibly due to increased cognitive load or distraction.

Recent studies have further examined these issues. Mahdi et al. (2024) conducted a meta-analysis that highlighted the effectiveness of multimedia glosses, particularly in enhancing recognition test performance. The study found that using a single type of gloss alongside content was more beneficial than employing multiple types of glosses. In another study, Zhang and Ma (2024) analyzed the overall impact of textual glosses and various glossing types—such as language, location, and mode—on second language (L2) vocabulary acquisition. They also examined how these factors moderated the variability in effect sizes. They concluded that the location and mode of glossing significantly influence the effectiveness of textual glosses on L2 vocabulary learning, whereas the language of the gloss does not have a clear impact. The outcome measure emerged as the only significant moderator variable. Furthermore, research by Rassaei and Folse (2024) explored the effects of word-level glosses in both the first language (L1) and second language (L2), as well as sentence-level L2 glosses, on L2 vocabulary acquisition. In their study, learners in the experimental groups could hover their computer cursor over underlined target words to view a gloss. The findings underscore the value of incorporating sentence-level glosses in instructional materials designed for L2 vocabulary teaching.

Given the diverse findings across studies, a meta-analysis could significantly contribute to our understanding of the overall impact of text-plus-picture glosses on incidental vocabulary learning. Such an analysis would not only synthesize existing research to provide a clearer picture of their effectiveness but also help identify key moderating factors. These factors might include learner characteristics, such as age and proficiency level, the design of the glosses, and various contextual variables that could affect learning outcomes. By examining these elements, a meta-analysis, like the present study, could guide educators and researchers in optimizing glossing strategies to enhance vocabulary acquisition in digital learning environments.

Literature Review

Multimedia Glossing and Incidental Vocabulary Learning

A gloss is defined as “a brief explanation of a word or phrase in a text” (Webb & Nation, 2017). Glosses provide learners with the lexical meaning of unfamiliar words, helping them understand texts without resorting to random guessing. It is widely recognized that texts with more than

5% unfamiliar words can hinder reading comprehension (Nation, 2013), as insufficient context makes successful guessing difficult. Glosses help bridge vocabulary gaps and enhance reading comprehension.

One area of gloss research that has garnered significant attention is gloss modality, which includes textual, pictorial, and auditory modes. These are often referred to as multimedia glosses (e.g., Ramazanali & Faez, 2019) or hypertext glosses (e.g., Yun, 2011). The current study focuses on multimedia glosses. Advances in computer technology have facilitated the development of electronic or hypertext glosses within computer-assisted language learning (CALL) environments (Abraham, 2008). A seminal study in this field was conducted by Chun and Plass (1996), who explored the effects of different gloss modalities: text definitions, text + picture glosses, and text + video glosses. Their findings indicated that text + picture glosses were the most beneficial for both productive and receptive vocabulary knowledge. This result was corroborated by Kost et al. (1999), who found similar benefits with text + picture glosses compared to text-only or picture-only glosses. Other studies (Al-Seghayer, 2001; Akbulut, 2007; Yoshii, 2006; Yoshii & Flaitz, 2002; Boers et al., 2017) also supported the superior effects of text + picture glosses. Teng (2023) supported the use of textual definition+video+word information on vocabulary learning and retention (see also Teng & Zhang, 2023, 2024). These findings align with the Theory of Multimedia Learning (Meyer, 2001). Boers et al. (2017) used eye-tracking technology to assess learners' attention to glosses, showing that learners spent more time on text + picture glosses, which led to better vocabulary retention. This supports the Noticing Hypothesis (Schmidt, 2001), which suggests that increased attention enhances retention.

However, some studies have reached different conclusions. Acha (2009) found that glosses with only words outperformed other modalities in both immediate and delayed tests (two weeks post-treatment). The author attributed these results to the low cognitive load on children due to their limited working memory capacity (Baddeley, 1997; 1999; 2012). Reanalyzing previous studies, Boers et al. (2017) questioned the effectiveness of adding pictures to glosses. Their results showed no significant differences between groups from different countries. Zou and Teng (2023) supported that pictures, GIFs were significantly more effective than video annotations, while text annotations were the least effective. They referenced Barcroft's (2015) view that pictures might distract learners from focusing on form. An eye-movement study by Warren et al. (2018) found that readers spent the least time on picture glosses but performed best in form recall and meaning recognition tests. This finding contradicts Boers et al. (2017), who suggested more attention leads to better retention. It also highlights the Depth of Processing theory (Craik & Lockhart, 1972), which posits that attention and memory are limited under conditions of multiple information inputs.

In summary, research on gloss modality effects is inconclusive, with disagreements not only in results but also in theoretical interpretations. Some argue against adding pictures, suggesting multimedia information may inhibit attention and memory, while proponents claim it frees attention from irrelevant details. Therefore, a meta-analysis could provide a comprehensive understanding of these effects.

Meta-Analyses on Incidental Vocabulary Learning and Glossing

One influential study by Abraham (2008) examined the differing impacts of paper glosses and computer-assisted glosses on vocabulary acquisition and reading comprehension. This meta-analysis included 11 studies and found that computer-assisted glosses had substantial effect sizes on vocabulary learning, with immediate posttest results averaging 1.40 and delayed posttest results averaging 1.25. Notably, the effect sizes for receptive assessments were significantly larger than those for productive assessments.

Another comprehensive meta-analysis by Yanagisawa, Webb, and Uchihara (2020) focused on various gloss formats. This study analyzed different types of glosses, including interactive, marginal, in-text, and glossary formats, as well as gloss language (L1 translations versus L2 definitions or synonyms) and gloss mode (textual, pictorial, and audio). The meta-analysis incorporated 42 studies, yielding 359 effect sizes, and revealed that glossing could lead to a 45.3% vocabulary acquisition rate compared to non-gloss conditions. However, no significant differences were found among the three gloss modalities.

Yun (2011) conducted another relevant meta-analysis investigating the effects of adding pictures to glosses on vocabulary learning. This study reported a medium effect size of $g = 0.46$ from 10 studies involving 1,560 participants. It found that learner proficiency and vocabulary test type significantly influenced the results. Despite these findings, some studies with opposing results, such as those by Acha (2009) and Boers et al. (2017), were not included, as noted by Boers et al. (2018). Therefore, it is essential to combine these dissenting studies with supportive ones to better understand the true effects and to explore additional moderating variables for a more comprehensive picture of this issue. In the study by Ramezanali et al. (2021), 22 studies were analyzed, yielding 26 independent effect sizes. The researchers examined 11 moderator variables, which included aspects such as the quality of the data sample, learner characteristics, gloss features, text attributes, and methodological elements. The findings revealed that the overall impact of adding an extra gloss mode was moderate ($g = 0.46$) for immediate posttests and small ($g = 0.28$) for delayed posttests. However, further analysis of the moderator variables showed that the effectiveness of additional gloss modes varies depending on several factors, including learner proficiency, the language of the gloss, the type of text (narrative versus expository), and the research design, such as the format of the test.

Recent meta-analyses have explored various aspects of glossing in language learning. In a study by Mahdi et al. (2024), the researchers highlighted the significant role of multimedia glosses in improving performance on recognition tests. They found that using a single type of gloss was more effective than employing multiple types simultaneously. Zhang and Ma (2024) conducted another study focusing on the impact of textual glosses and different glossing characteristics, such as language, position, and mode, on second language (L2) vocabulary acquisition. After analyzing 20 studies, they concluded that textual glossing had a moderate effect on L2 vocabulary acquisition, with its benefits enduring over time. They also discovered that the position and mode of glossing significantly influenced its effectiveness, whereas the language used for glossing did not. Among the factors examined, the outcome measure was identified as the only significant moderator affecting vocabulary acquisition. Kim et al. (2024) investigated the comparative effects of L1 and L2 glossing on L2 reading comprehension and vocabulary learning. By analyzing 78 effect sizes from 26 studies, they found that L1 glossing was generally more effective than L2 glossing (Hedge's $g = .33$, $SE = .09$, $p < .001$), especially in immediate vocabulary posttests. The study also revealed that learners' L2 proficiency level significantly affected glossing effectiveness, with L1 glossing being particularly beneficial for beginners. In another comprehensive analysis, Mohsen et al. (2024) combined meta-analytic methods with document co-citation analysis to explore the thematic connections in studies about glosses and L2 vocabulary learning. Their review of 18 studies involving 2,991 participants showed that glosses had a substantial overall effect on L2 vocabulary acquisition (Hedge's $g = 1.27$). However, the effect size varied based on learner proficiency, being larger for recognition tests and moderate for vocabulary recall tests.

Given these insights, there is a clear need for another comprehensive meta-analysis that not only synthesizes existing findings but also incorporates a broader range of studies. Such an analysis should aim to address the limitations of previous research, including small sample

sizes and limited contextual diversity. Additionally, it should consider emerging digital and interactive glossing technologies and their potential to enhance vocabulary acquisition across different learner profiles and educational settings. The present study provides a more nuanced understanding of how glosses can be optimized to support diverse learning needs.

Present Study

To sum up, the above studies on meta-analysis on glossing result in a gap of effect size and neither of them include some opposite studies in their analyses. Moreover, the effects of what moderating variables are still unclear. Therefore, more work is needed to better understand this area. To clarify the dispute, the present meta-analysis study on the effects of multimedia glosses on incidental vocabulary learning was conducted. The present study is guided by the following research questions:

1. What are the overall effects of multimedia glosses (text + picture) on L2 incidental vocabulary learning compared to the single media gloss (text gloss)?
2. What are the possible moderating factors that could affect the effects of multimedia glosses on L2 incidental vocabulary learning?

Methodology

Literature Search

After reviewing previous studies in this area, a comprehensive literature search was conducted to include as many relevant studies as possible. The search utilized keywords such as “gloss” (including variations like glosses, glossing, glossary using the search term gloss*), and “annotation” (noting that in computer science, annotations refer to different concepts, so they were paired with terms like vocabulary learning/acquisition). Additionally, terms such as “incidental vocabulary learning” and “incidental vocabulary acquisition” were used. These terms, when searched in conjunction with gloss or annotations, were drawn from previous meta-analyses (Abraham, 2008; Kim et al., 2024; Yun, 2011).

The databases used for this search were selected based on recommendations by Plonsky and Oswald (2015) for linguistics and applied linguistics research. These included the Educational Resources Information Center (ERIC), Linguistics and Language Behavior Abstracts (LLBA), and PsycINFO. Other sources included JSTOR and ProQuest—Dissertations & Theses Global A&I: The Humanities and Social Sciences Collection, which are advised by other researchers to minimize publication bias in meta-analyses. To include studies from China, the Chinese database CNKI was also utilized.

The search strategy involved keyword searches and cross-referencing, assisted by Google Scholar and Web of Science, focusing on studies that cited the influential paper by Chun & Plass (1996). The search yielded 119 studies, including 7 theses and 12 studies from China. The main journals where these papers were published include *CALICO Journal*, *Studies in Second Language Acquisition*, *Language Learning*, *The Modern Language Journal*, *TESOL quarterly*, *System*, *Language Learning & Technology*, *Reading in a Foreign Language*, *ReCALL*, and *Computer Assisted Language Learning*.

Inclusion and Exclusion Criteria

To achieve the objectives of this meta-analysis, the studies identified in the search needed to be filtered according to specific criteria based on the definitions of incidental vocabulary learning

and glosses, as well as the research questions. The criteria for selecting eligible studies were as follows:

Gloss Conditions: The studies must include pictorial glosses at a minimum. Studies that compared different languages used or compared multiple-choice glosses with pre-given glosses were excluded.

Comparison of Gloss Conditions: There should be a clear comparison among different gloss conditions, rather than a mixture of glosses. For example, Abraham (2007) used multimedia glosses, but his study focused on comparing choice look-up with forced look-up, with all glosses available to both groups. Such studies were excluded.

Dependent Variables: The studies should include vocabulary retention or recall as dependent variables. While some studies might report both reading comprehension scores and vocabulary test scores, only vocabulary scores were considered. Studies reporting only reading scores were excluded.

Incidental Nature of Vocabulary Tests: To ensure the incidental nature of vocabulary learning, participants should not be informed in advance about vocabulary tests, aligning with Hulstijn's (2013) definition. However, participants could be informed of a reading comprehension test to focus their attention on understanding the text.

Effect Size Calculation: Studies should report information necessary for calculating effect sizes. For between-group experiments, means and standard deviations or t-values are required. For within-group experiments, means and standard deviations of pretest and posttest, or paired t-values between pretest and posttest, should be available. Some studies may directly report d-values or correlations.

Input Type: The input for vocabulary learning should be written rather than spoken. Spoken input should use captions to provide information about unfamiliar words instead of glosses.

Two raters selected the eligible papers: one was the author of this study, and the other was a graduate student majoring in linguistics and applied linguistics. The first rater identified 14 studies, while the second identified 16, with eight studies overlapping. The interrater reliability, measured using Cohen's Kappa, was 0.46, indicating a moderate level of agreement. To enhance the reliability of the results, the raters discussed controversial studies and adjusted some criteria, notably adapting the second criterion to compare text gloss with text plus picture gloss.

Ultimately, 22 studies meeting the six criteria were selected and are listed in references. Studies that were excluded, along with reasons for exclusion, are listed in Appendix A, following the recommendation by Norris and Ortega (2006) to provide transparency for other researchers. While de Vos et al. (2018) included all excluded studies in their appendix, this was not feasible for the current study.

A common issue in meta-analyses is the discrepancy between study quality and publication bias. Study quality is defined as the "fit between a study's goals and the study's design and implementation characteristics" (Valentine, 2009, pp. 130). High-quality studies should align their design and implementation with their research questions, an important criterion in peer-reviewed journals. However, if only significant studies are published, it could lead to

publication bias, invalidating the meta-analysis results. To address this, dissertations and theses were included in the search to ensure both study quality and minimize publication bias. However, no dissertations or theses met the inclusion criteria.

Coding Scheme

Between the inclusion of studies and the actual meta-analysis lies the coding procedure. As Plonsky and Oswald (2015) aptly put it, “If meta-analytic results are the meal that readers will feast upon, coding is what happens back in the kitchen to prepare the food for the meal” (pp. 111). This meta-analysis utilized the coding scheme from De Vos et al. (2018) due to its clarity and comprehensibility. Additionally, for specific variable categories, Abraham’s (2008) version was preferred to facilitate easier coding. The revised coding scheme is presented in Table 1.

To ensure feasibility and adapt to the specific characteristics of the filtered studies, some modifications were made. For instance, since this analysis required a comparison between text gloss and text + picture gloss groups, the comparison group variable was removed, whereas De Vos et al. (2018) included it in their study. Another removed variable was publication type, as all studies in this analysis were sourced from journals. New variables added include the use of distractors and non-words, which can be considered moderating variables, following the approach of Uchihara et al. (2019).

Coding

The coding scheme serves for the coding process. Two coders did the coding independently. One is the author of this study while another is a postgraduate majoring in foreign linguistics and applied linguistics. Due to the second coder’s unfamiliarity of the glossing, before the coding, the first coder gave a coding training including the general design of gloss studies, the structure of the coding scheme and the attention needed when coding. The interrater reliability was 0.78 and the disagreement was resolved through discussion.

The 22 studies yielded 84 effect sizes because some studies included three groups (like Chun & Plass, 1996) or there were several vocabulary tests (like both meaning recognition and meaning recall were included) or delayed vocabulary tests were employed. The detailed information of 22 studies is presented in the Appendix B. Here, the narrative description of the study characteristics is shown with some visual tables and figures to present. Among them, two studies were published in the 1990s, nine in the 2000s, six in the 2010s and five in the 2020s. Figure 1 shows the distribution of the year of publication.

The age group is divided into four groups: elementary school, middle school, mixed group and university. The result of age groups coding is shown in Table 2.

Based on Table 2 above, there is a huge imbalance between different age groups (with university thirteen times than the other groups). According to the previous meta-analyses, the imbalanced distribution of subgroups would overestimate the difference between different groups. Therefore, this variable was not included. Among the 22 studies, half of those reported the age means or age range of the participants, which poses the difficulty to the data analysis. There were 2,829 participants in total. However, only the group sizes of text gloss and the text + picture gloss were needed for the calculation of effect sizes. Then, the L2s studied in the 22 studies included four languages: English, French, German and Spanish. The description of L2s is shown in the Table 3.

The same conclusion could be drawn from the Table 2 as age group. The L2s studied were distributed in an unfair way and the main language to study is English. As for the L2 proficiency variable, some studies reported the proficiency level like beginners, intermedia and advance while some just reported the years of learning L2.

Table 1 Coding scheme of this study.

Study characteristics	No.	The number of each study
	Author	The name of author (s)
	Year	The publication year
	Appeared in	The journal source
	Age group	University/Middle school/Elementary school
	Age	Report the mean age or age range of participants
	Group size	Report the whole group size
	L1	Report the L1 background
	L2	Report the L2 studied
	L2 proficiency	Report the L2 proficiency
Effect size characteristics	No.	The number of each study
	Author	The name of author (s)
	Year	The publication year
	Gloss type	Text/Text + picture
	Gloss language	L1/L2
	Number of words glossed	Report the number of glossed words
	Non-words used	Yes/No
	Testing instrument	Report what testing instrument was used
	Vocabulary test	Receptive/Productive
	Distractor used	Yes/No
	Delayed	Yes/No
	Timing	Report the timing between immediate and delayed test
	Group size	Report the size of each group
	mean(pre)	Report the means of pretest
	mean(post)	Report the means of immediate posttest
	mean(delay)	Report the means of delayed posttest
	SD (pre)	Report the standard deviations of pretest
	SD (post)	Report the standard deviations of immediate posttest
	SD (delay)	Report the standard deviations of delayed posttest
	t-value	Report the t-value if the means and SD are missed
	chi-square	Report the chi-square value if the means and SD are missed

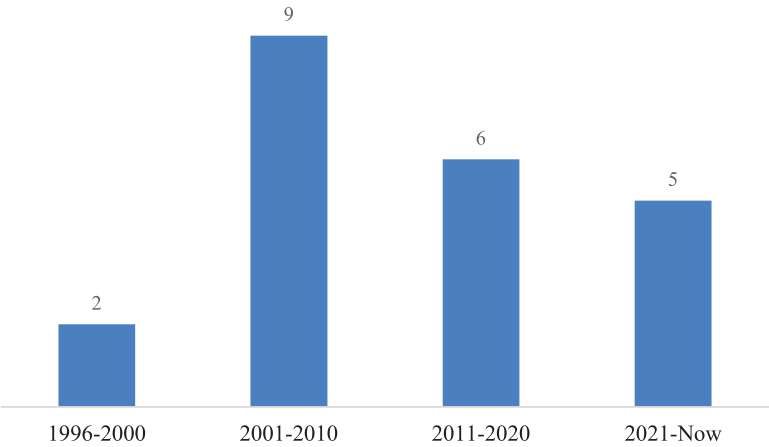


Figure 1 Year of publication.

Table 2 Count of age groups.

Age group	Count
Elementary school	1
Middle school	1
Mixed group	1
University	19

Table 3 Calculation of L2 studied.

L2	Count
English	17
French	1
German	2
Spanish	2

The moderating variables were mostly included in the effect size characteristics: gloss language, number of words glossed, testing, non-words used, and distractor used. Among them, the gloss language (Taylor, 2009; 2013), non-words used (Uchihara et al., 2019), and testing (Abraham, 2008; Yun, 2011) have been studied but yielded disagreeable results. The number of words glossed was mentioned in Abraham’s study (2008). We included as a moderating factor as well.

Data Analysis

The collected data were analyzed using the Comprehensive Meta-Analysis (CMA) software version 3.0, a widely used tool in meta-analyses. This software can integrate various research

designs, including between-group experiments and quasi-experiments, to calculate the effect sizes of the included studies. It also allows for comparisons of effect sizes within different sub-groups using the Q-test for heterogeneity, which will be discussed later. Additionally, CMA can compute publication bias and perform meta-regression analyses. As previously mentioned in section 4.3, publication bias is assessed to determine if all eligible studies show positive results, as studies published in peer-reviewed journals are often statistically significant, potentially skewing meta-analysis results. Meta-regression examines the correlations between moderator variables and effect sizes.

Firstly, the effect sizes of different designs were calculated using Hedge's g , chosen because it reduces the overestimation of standard errors in small samples compared to Cohen's d (Borenstein, 2009). It is important to note that within a study, there may be sub-groups (such as the three groups in Boers et al.'s 2017 study) or different test timings, like immediate and delayed tests. Effect sizes for these sub-groups and timings were calculated separately, and heterogeneity tests were used to compare effect sizes across different groups. Different studies might report data differently; some provide sample sizes, means, and standard deviations, while others only report t -values or chi-square values. CMA can handle these variations if the necessary data are provided. The formulas used for calculations are not detailed here, as the software performs these automatically.

After calculating the effect sizes for each study, a random-effects model was employed to aggregate effect size estimates across studies. This model assumes that effect sizes vary across studies, unlike a fixed-effects model, which assumes a single effect size with differences due only to sampling errors (Shadish & Haddock, 2009). Considering the diverse methods and participants in this meta-analysis, a random-effects model was more appropriate. In addition to reporting weighted effect sizes, CMA provides results of heterogeneity tests and confidence intervals to illustrate variations among studies. Comparisons among subgroups or moderator variables also use heterogeneity tests to show variations.

Simultaneously, publication bias and meta-regression were assessed. Publication bias examines whether the selected studies are all positive or statistically significant. Rosenthal's Fail-safe N , Orwin's Fail-safe N , and the Trim and Fill model were used to evaluate publication bias in this meta-analysis. Meta-regression focuses on the coefficients of moderator variables, an aspect often overlooked in previous meta-analyses. Detailed descriptions of publication bias and meta-regression will be presented in the results section.

Results

Overall Effects of Multimedia Glosses on L2 Incidental Vocabulary Learning

To address the first research question, a meta-analysis was conducted to compare the effects of multimedia glosses (text + picture) with single text glosses. This analysis included 22 primary studies and employed a random-effects model using CMA software. The results indicated a medium effect size ($g = .727$), according to Plonsky and Oswald's classification (2014), where an effect size around .4 is considered small, .7 is medium, and 1.0 is large. Detailed effect sizes and confidence intervals for each study are illustrated in Appendix C. It is important to note that some studies reported different outcomes, such as meaning recall versus meaning recognition, or investigated delayed retention of vocabulary, leading to multiple outcomes from a single study, as detailed in Appendix C. Additionally, Table 4 presents descriptive statistics for the overall effect size and the heterogeneity test (Q-test). For publication requirements, Appendix C includes the effect size and confidence interval for all 84 effect sizes.

Aside from the effect size (g), Table 4 shows that the confidence interval ranges from .544 to .910, further indicating a bigger effect. The heterogeneity test examines the variance among the 84 effect sizes using the Z-value and Q-test. The Z-value revealed significant differences within these effect sizes ($p < .001$). However, the Z-value can sometimes overestimate sampling errors, making the Q-test a more reliable method for assessing heterogeneity, as it functions similarly to a chi-square test on the distribution of effect sizes around zero (Borenstein, 2009). In this study, the Q value was 942.953 ($p < .001$), significantly exceeding the degrees of freedom ($df=83$), indicating heterogeneity across studies. This finding rejects the hypothesis that variance across studies is zero. Furthermore, the I^2 value of 91.198% suggests that heterogeneity accounts for 91% of the variation in treatment effects (Higgins & Thompson, 2002). This result implies that, in addition to sampling errors assumed by the fixed-effects model, other factors influence variations across studies, leading to the second research question.

At the same time, Figure 2 shows the funnel plot of publication bias, which displays the relationship between SD (standard errors) and effect size. Previous funnel plot is a relationship

Table 4 Descriptive statistics of overall effect sizes

Model	Effect size and 95% confidence interval					Test of null (2-Tail)	Heterogeneity test		
	N	g	SE	Lower limit	Upper limit	Z-value	Q-value	df (Q)	I-squared
Fixed	84	.557	.027	.503	.611	20.346***	942.953***	83	91.198
Random effects	84	.727	.093	.544	.910	7.784***			

Note: SE = standard error; Q value = heterogeneity of variance test; *** $p < 0.001$.

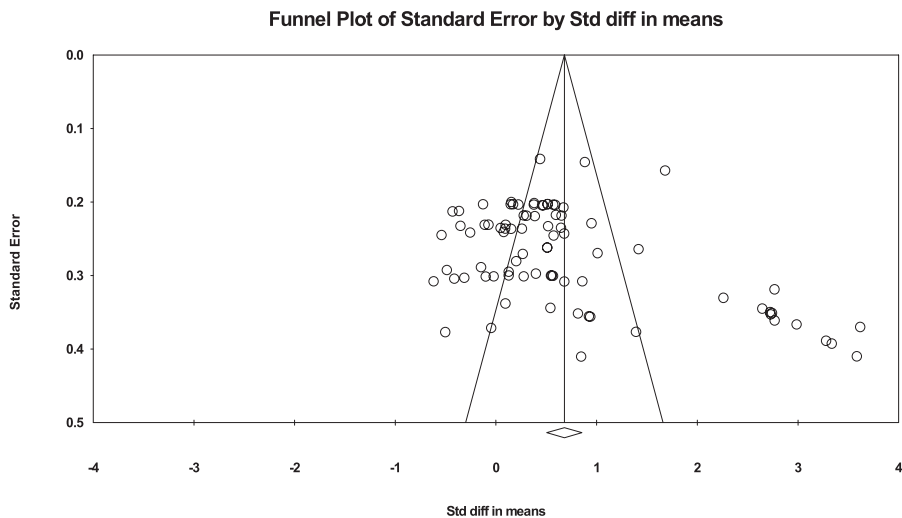


Figure 2 Funnel plot of publication bias of 84 effect sizes.

between study size and effect size, which was considered to be difficult to find asymmetry (Borenstein, 2009). From Figure 2, the distribution of effect sizes is symmetry. The diamond in the bottom of the figure represents the confidence interval of effect size and the central line represents the weighted effect size. The effect sizes are distributed around the central line.

The results presented above reflect the overall findings from 84 effect sizes, which were analyzed separately rather than combined. As Raudenbush (2009) noted, outcomes from the same sample groups, such as immediate and delayed tests, are often related, and treating them as separate can introduce errors into statistical analyses. Consequently, an analysis combining the results of immediate and delayed tests was also conducted. Previous meta-analyses on glosses (Abraham, 2008; Yun, 2011) typically distinguished between immediate and delayed tests, using the Q-test to compare their effects. However, these analyses suggested that the tests were unrelated. In contrast, this study considers immediate and delayed tests as related because they were conducted on the same sample, and thus, they were not included in the heterogeneity test (Q-test) for comparison. On the other hand, results from different types of tests, such as meaning recognition and meaning recall, were not combined, as they involve different aspects of vocabulary knowledge (Nation, 2013; Read, 2000). This approach resulted in a reduced number of effect sizes (22) and an increase in their values. Detailed effect sizes for each test are provided in Appendix C.

Additionally, Table 5 presents the descriptive statistics for the overall effect size, confidence interval, and heterogeneity test. As shown in Appendix C and Table 5, the revised effect size is 0.882, which is bigger than the original. The null hypothesis test and the heterogeneity test remain significant, indicating that variance across studies is still substantial. Therefore, further research is needed to explore these variances.

The research design itself can also influence the effects observed. For instance, vocabulary retention and recognition demand different cognitive processes: retention involves long-term memory, while recognition primarily relies on working or short-term memory.

Effects of Moderating Factors on Multimedia Glosses

This section explores which factors contribute to this variance. Five potential moderating factors are considered: gloss language, number of words glossed, use of non-words, testing formats, and the presence of distractors.

Heterogeneity Test of Categorical Factors

Heterogeneity tests assess the variation between or across different subgroups or factors. Each of the five categorical factors was analyzed separately. The Q-test was employed to compare the

Table 5 Descriptive statistics of 22 effect sizes with combination of immediate tests and delayed tests.

Model	Effect size and 95% confidence interval					Test of null (2-Tail)	Heterogeneity		
	N	g	SE	Lower limit	Upper limit	Z-value	Q-value	df (Q)	I-squared
Fixed	22	.557	.027	.503	.611	20.346***	781.717***	21	97.314
Random	22	.882	.173	.544	1.220	5.114***			

Note: SE = standard error; Q value = heterogeneity of variance test; *** p < 0.001.

effects of different subgroups but was not used for continuous variables, such as the number of words glossed. Continuous variables were analyzed using meta-regression alongside categorical variables.

The first factor examined is the language used in glosses, prioritized because it directly relates to the gloss itself, unlike other variables linked to design features such as testing. To facilitate comparisons, languages were coded as L1 (first language) or L2 (second language) rather than specific languages like English or French. Table 6 presents the descriptive statistics of language use across 84 studies. Of these, 20 studies used the learners’ first language for glossing target words, while 62 used a second language and only 2 studies used both L1 and L2 glosses. Although a random-effects model was initially used due to the complexity of research designs, a fixed-effects model was applied here to combine subgroup effects into an overall effect (Borenstein, 2009). Table 6 indicates that studies using L2 glosses ($g = .674$) produced much larger effects than those using L1 glosses ($g = .276$) and a little larger effect than L1+L2 glosses ($g = .642$). However, since the number of using L1+L2 glosses is rather limited, the interpretation of this effect should be cautious. A significant difference was found among glosses languages ($Q_{\text{between}} = .43.418, p < .001$). The result indicates that when using text+picture gloss, the text presented with L2 will be beneficial for vocabulary learning.

The second factor analyzed is vocabulary testing, which is divided into receptive and productive vocabulary tests. The analysis revealed that 43 studies used productive tests, while 41 used receptive tests. A fixed-effects model was applied to analyze the overall effects and compare differences between the groups. Table 7 presents the Q-test results for the two types of testing. It indicates that studies using receptive tests ($g = .615$) produced a larger effect than those using productive tests ($g = .506$), with a significant difference between the two groups ($Q = 4.008, p < 0.05$). Notably, the Q-test for studies employing both receptive tests and productive ones indicated a significant level, suggesting variance exists across the 84 studies. In other words, studies using different tests tend to produce different effects due to sampling or design features.

Before moving on to the next part of the discussion, it is important to note that the spacing of tests, such as immediate and delayed tests, often garners researchers’ attention.

The next variable under consideration is the use of distractors in vocabulary tests, with results shown in Table 8. 20 studies incorporated distractors, while 64 did not. The table reveals that studies using distractors had lower effects, but not reached a significant level ($Q_{\text{between}} = 1.822, p > .05$). This result is not beyond expectation since using distractors would make the test results more accurate. Meanwhile, the Q-value for studies not using distractors showed significant

Table 6 Q-test of gloss language variable.

Group	Number of Effect Sizes	g	Lower Confidence	Upper Confidence	Z-Value	Q-value
L1	20	.276	.177	.375	5.442***	65.755***
L2	62	.674	.609	.738	20.408***	830.005***
L1+L2	2	.642	.267	1.017	3.359***	3.775
Total between						43.418***

Note: L1 = Glosses with first language; L2 = Second language; *** $p < .001$.

Table 7 Q-test of the testing variable.

Group	Number of Effect Sizes	g	Lower Confidence	Upper Confidence	Z-Value	Q-value
P	43	.506	.432	.579	13.470***	547.513***
R	41	.615	.537	.694	15.379***	391.432***
Total between						4.008*

Note: P = Productive test; R = Receptive test; *** $p < .001$; * $p < .05$.

Table 8 Q-test of distractor use variable.

Group	Number of Effect Sizes	g	Lower Confidence	Upper Confidence	Z-Value	Q-value
0	64	.574	.515	.633	19.019***	911.013***
1	20	.477	.350	.605	7.353***	30.119
Total between						1.822

Note: 0 = no distractor used; 1 = distractor used; *** $p < .001$.

Table 9 Q-test of nonword use variable.

Group	Number of Effect Sizes	g	Lower Confidence	Upper Confidence	Z-Value	Q-value
0	77	.598	.543	.654	21.229***	888.583***
1	7	-.246	-.489	-.003	-1.988*	10.140
Total between						44.228***

Note: 0 = authentic words; 1 = nonword use; *** $p < .001$; * $p < .05$.

differences across studies ($Q = 911.013$, $p < .001$), indicating inconsistent effects due to other testing variables. Interestingly, there exists no significant difference between studies using distractors ($Q = 30.119$, $p = .05$), suggesting a rather similar effect within studies. The I^2 value for this variable is 36.919%, as a small effect suggests that the variance is due to sampling errors rather than other factors (Higgins, & Thompson, 2002). This result underscores the importance of using distractors in vocabulary tests. Despite its significance, this variable has not received adequate attention, and its effects warrant further investigation.

The final categorical variable is the use of non-words, also examined by Uchihara et al. (2019). Among the 84 studies, only 7 used non-words as target words, with results shown in Table 9.

Studies using non-words yielded a negative effect ($g = -.246$), indicating adverse effects from their use. The heterogeneity test revealed a significant difference between the two subgroups ($Q_{\text{between}} = 44.428, p < .001$). However, this result should be interpreted cautiously due to the large disparity in subgroup sizes, which can overestimate heterogeneity effects, as noted in meta-analysis guidelines (Borenstein, 2009).

Meta-Regression of Moderating Variables

In the current study, the continuous variable “number of words glossed” could not be analyzed using the Q-test, as this test is not suitable for assessing the effects of continuous variables. Instead, meta-regression, a more appropriate model, was employed to integrate both categorical and continuous variables.

Table 10 presents the meta-regression model, including the coefficients, confidence intervals (CI), and hypothesis test results (Z-value and p-value). This table specifically highlights one condition within each categorical subgroup: L2 gloss, receptive testing, distractor use, and non-word use. The coefficients for other conditions are the same but with opposite signs. The analysis employed a random-effects model, which assumes variability between studies, as confirmed by the $Q\text{-value} = 24.01, p < .001$.

From the table, it is evident that only the nonword variable reached a significant level, with a coefficient of -1.344 . This indicates that increased use of nonwords is associated with a decreased effect size. This finding aligns with the heterogeneity test results, which also showed that nonword use negatively impacts the effectiveness of multimedia glosses. Another significant variable is the gloss language with $Q = 11.69, p < .01$, which means that different language used in text gloss may lead to different effects on vocabulary learning and using L2 is more beneficial than other types.

Apart from these two variables, none of the others achieved significance, which contrasts with the heterogeneity test results. However, these two methods offer different insights. Meta-regression focuses on the relationships between variables and effect size, while the heterogeneity test compares effect sizes across different groups. Meta-regression examines between-variable effects, whereas the heterogeneity test looks at within-variable effects. By analyzing the data from these different angles, the two methods together provide a more comprehensive understanding of the research questions.

Table 10 Meta-regression of five factors.

Covariate	Coefficient	95% Lower	95% Upper	Z-value	p
Intercept	.403	-.136	.942	1.47	.142
Words	-.006	-.027	.016	-.53	.597
GL: L2	.794	.338	1.250	3.41	< .001
GL: L1+L2	.695	-1.008	2.399	.80	.424
T: R	.040	-.354	.435	0.20	.842
D: Y	-.248	-.723	.228	-1.02	.307
Nonword: Y	-1.344	-2.113	-.575	-3.42	< .001

Note: GL= Gloss languages; T: R = Tests: Receptive; D: Y = Distractors: Yes; Nonwords: Y = Nonwords: Yes.

Discussions

Overall Effect of Multimedia Gloss on Incidental Vocabulary Acquisition

In comparison to previous studies, this meta-analysis reveals a relatively medium effect size for the use of multimedia glosses (text + picture gloss), with effect sizes of $g = .727$ across 84 effect sizes and $g = .882$ across 22 studies (Ramenazali et al., 2021; Yun, 2011). A detailed examination of the studies included suggests that earlier research often incorporated fewer studies (like Yun, 2010) or some studies published in non-SSCI indexed journal (like Ramezanali et al., 2021), which may have led to an underestimation of the weighted effect sizes.

In Mohsen et al. (2024), multimedia glosses demonstrated a substantial effect size on L2 vocabulary acquisition (Hedge's $g = 1.27$). This suggests that the effect of picture plus text gloss is nearly equivalent to that of text-only gloss on incidental vocabulary learning. To some extent, this supports Mayer's cognitive theory of multimedia learning (2009), which posits that presenting both verbal and pictorial information enhances learning more than text alone. Learners can process dual code channels—text and picture—simultaneously, engaging more processing procedures. This is evidenced by Warren et al.'s eye-tracking study (2018), where participants spent more time on picture glosses. However, Mohsen et al. (2024) attempted to summarize several meta-analyses on glossing rather than solely focusing on the text + picture gloss. For this meta-analysis, the effect of using text + picture gloss is much better than text-only gloss, which lends support to studies above.

According to the working memory theory (Baddeley, 1999; Baddeley, 2012), which distinguishes between the phonological loop and visuo-spatial sketchpad, humans process verbal and pictorial information separately. When encountering a word, people tend to visualize it, but an illustrated picture can simplify this process, aiding quicker information processing.

However, the anticipated large effect of multimedia glosses can provide support to the theory of multimedia learning. The result aligns with its principles, which emphasize that presented information should be coherent to activate cognitive processing capabilities like selection and organization (Mayer, 2009).

Moderation Factors

The analysis reveals that the language used in glosses does moderate the effectiveness of multimedia glosses, a finding that aligns with Yoshii's (2006), which identified an interactive effect between gloss language and gloss mode. Yoshii (2006) found significant differences when comparing multimedia glosses in L1 and L2 to text-only glosses. However, this result contradicted with the meta-analysis by Ramezanali et al. (2021) and Mahdi et al. (2024), which concluded that glossing language does not have a moderating impact on the effects of multimedia-glosses. One of the reasons for difference can be accounted by numbers of analysis. In Mahdi et al.'s study, there only 5 studies involved in the analysis of this variable. However, this study echoed with Ramezanali et al.'s findings (2021) that using L2 glosses is more effective. The discrepancy in findings could also be attributed to Nation's (2013) assertion that L2 glosses are more beneficial for learners with extensive vocabulary knowledge. Given that most studies in this meta-analysis focus on university students, who typically possess a large vocabulary, the language of the gloss may not significantly influence outcomes.

Multimedia glosses appear to be more advantageous for receptive vocabulary knowledge than for productive knowledge, aligning with previous studies (Teng, 2023; Teng & Zhang, 2023; Teng & Zhang, 2024). This is unsurprising, as it is widely accepted that receptive knowledge is

easier to acquire (Nation, 2013; Schmitt, 2008). Additionally, incidental vocabulary learning through reading is primarily a receptive process, requiring learners to recognize word meanings without needing to use them in context. From the perspective of multimedia learning theory (Mayer, 2009), this process involves memorization and rote learning rather than understanding and meaningful learning. This finding is consistent with the conclusions of Yun (2011) and Abraham (2008), who noted that the type of vocabulary test used can influence gloss outcomes. Furthermore, it aligns with De Vos et al. (2018), who found significant differences in incidental vocabulary learning when tested receptively versus productively. Supporting Ramezanali et al. (2021), analyses of moderator variables show that the effectiveness of additional gloss modes is influenced by factors such as learner proficiency, gloss language, text type (narrative vs. expository), and research design (e.g., test format). However, other studies often compared different testing effects without revealing the distinct impacts of each test type, as the present study does, indicating that receptive tests typically yield similar effects.

According to Nation (2013) and Read (2000), the use of distractors in vocabulary tests can affect results. Distractors in pretests can prevent participants from intentionally focusing on target words, while distractors in posttests can reduce random guessing. The moderating analysis of this distractors use support their claims. When using distractors, the effect size is smaller than not using it, which could lead to more accuracy interpretation of vocabulary learning. Notably, some studies claim to use distractors in multiple-choice questions, but this complicates coding and is not regarded as true distractor usage.

The finding that non-words negatively affect outcomes contradicts Uchihara et al.'s (2019) study, which found a strong positive correlation between non-word use and incidental vocabulary learning ($r = .50$). However, their study examined the effects of repetition on incidental vocabulary learning, with non-words serving as a moderating variable. They suggested that unfamiliar items attract more attention. In this study, the focus was on glosses, and since lexical information was already provided, learners might concentrate more on the lexical information and picture aids, potentially explaining the negative effects, aligning with Ramezanali et al. (2021)'s findings. This could also be attributed to limited memory capacity (Baddeley, 1999), as target words can challenge learners, and excessive information may burden their processing, leading to negative effects. This result applies to authentic words, suggesting that providing multimedia glosses might not be necessary for low-level or complex words. The raw data indicate that only two studies used non-words (Boers et al., 2017; Boers et al., 2017), and despite differing results, only seven effect sizes involved non-words as target vocabulary. Moreover, the negative effects stemmed from one of these studies. Therefore, this result should be interpreted cautiously, and future research should explore whether non-words consistently lead to negative outcomes.

Limitations and Implications

Despite offering a comprehensive overview, this meta-analysis has certain limitations. Firstly, the number of studies analyzed is restricted; only 22 studies met the strict inclusion criteria, which is a relatively small sample size for a meta-analysis. Additionally, the factors examined do not fully account for the variance observed across studies. Due to constraints in time and resources, factors such as learner characteristics and the frequency of target words were not considered, though they could yield interesting insights. Moreover, the conflicting results from the heterogeneity test and meta-regression require further investigation and clarification. Notably, the analysis overlooks other gloss modes like text with audio or video, which are of significant interest in second language studies and computer-assisted language learning. However,

text-plus-picture glosses are easier to define, whereas other modes are challenging to compare due to some studies equating video with both audio and picture, complicating comparisons.

Based on the conclusions, the implications can be divided into research and teaching purposes. From a research perspective, it is suggested that picture aids may not significantly enhance incidental vocabulary learning, as indicated by the smaller effect size. It is recommended to expand study samples to include groups such as middle and primary school learners, who might prefer picture glosses due to their developing language skills. The popularity of picture books also supports further research on these groups. Additionally, within-group studies could compare vocabulary knowledge improvement, as vocabulary learning is cumulative, and glosses can help reinforce memory to some extent. Future studies should also explore learner characteristics, such as perceptual styles and proficiency, to better account for learner-related factors in the variance across studies.

The second point concerns the measurement of incidental vocabulary learning outcomes. While multimedia glosses can consolidate receptive knowledge, productive vocabulary knowledge may be neglected. Future research should focus on enhancing productive vocabulary through additional output tasks or activities that promote deeper processing of target words, rather than merely providing glosses. It is also suggested to use distractors to minimize random guessing in vocabulary assessments.

The third point addresses the ecological validity of using non-words in studies. While non-words are used in second language acquisition research to eliminate the influence of prior knowledge, they are less suitable for classroom-based learning. Glosses are more appropriate for classroom settings due to technological advancements, and using non-words may unnecessarily increase the learning burden.

Regarding teaching, multimedia glosses combining text and pictorial elements can aid students in building vocabulary, particularly receptive vocabulary knowledge. However, productive vocabulary knowledge might be better developed through activities like writing sentences with target words. Additionally, using both first and second languages in glosses is recommended for lower proficiency learners, while second language glosses may be more suitable for higher proficiency learners.

References (*studies included in the meta-analysis)

- Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, 21(3), 199–226. <https://doi.org/10.1080/09588220802090246>
- *Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. *British Journal of Educational Technology*, 40(1), 23–31. <http://doi.org/10.1111/j.1467-8535.2007.00800.x>
- *Akbulut, Y. (2007). Effects of multimedia annotations on incidental vocabulary learning and reading comprehension of advanced learners of English as a foreign language. *Instructional Science*, 35(6), 499–517. <http://doi.org/10.1007/s11251-007-9016-7>
- *Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology*, 5(1), 202–232. <http://dx.doi.org/10.125/25117>
- Baddeley, A. (1997). *Human memory: Theory and practice*. Psychology Press.
- Baddeley, A. D. (1999). *Essentials of human memory*. Psychology Press/Taylor & Francis.
- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63(1), 1–29. <https://doi.org/10.1146/annurev-psych-120710-100422>
- Barcroft, J. (2015). *Lexical input and vocabulary learning*. John Benjamins.
- *Boers, F., Warren P., Grimshaw G., & Siyanova-Chanturia A. (2017). On the benefits of multimodal annotations for vocabulary uptake from reading. *Computer Assisted Language Learning*, 30(7), 709–725. <https://doi.org/10.1080/09588221.2017.1356335>

- *Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System*, 66, 113–129. <https://doi.org/10.1016/j.system.2017.03.017>
- Borenstein, M. (2009). (Ed.). *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>
- *Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, 80(2), 183–198. <https://doi.org/10.1111/j.1540-4781.1996.tb01159.x>
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- de, Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning*, 68(4), 906–941. <https://doi.org/10.1111/lang.12296>
- *Farley, A., Ramonda, K., & Liu, X. (2012). The concreteness effect and the bilingual lexicon: The impact of visual stimuli attachment on meaning recall of abstract L2 words. *Language Teaching Research*, 16(4), 449–466. [10.1177/1362168812436910](https://doi.org/10.1177/1362168812436910)
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statist. Med.*, 21, 1539–1558. <https://doi.org/10.1002/sim.1186>
- Hustijn, J. H. (2001). Intentional and incidental second language vocabulary learning: a reappraisal of elaboration, rehearsal and automaticity. In P. Robinson. (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge University Press.
- Hustijn, J. H. (2003). Incidental and intentional learning. In C. Doughty, & M. Long. (Eds.), *The handbook of second language acquisition* (pp. 349–381). Blackwell Pub.
- Hustijn, J. H. (2013). Incidental learning in second language acquisition. In C. Chapelle. (Ed.), *The encyclopedia of applied linguistics volume V* (pp. 2632–2637). Blackwell.
- Hulstijn, J. H., Hollander, M. & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, 80, 327–339. <https://doi.org/10.1111/j.1540-4781.1996.tb01614.x>
- *Jones, L., & Plass, J. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations. *Modern Language Journal*, 86(4), 546–561. <https://doi.org/10.1111/1540-4781.00160>
- *Kim, D., & Gilman, D. (2008). Effects of text, audio, and graphic aids in multimedia instruction for vocabulary learning. *Educational Technology & Society*, 11(3), 114–126.
- Kim, H. S., Lee, J. H., & Lee, H. (2024). The relative effects of L1 and L2 glosses on L2 learning: A meta-analysis. *Language Teaching Research*, 28(1), 7–28. <https://doi.org/10.1177/1362168820981394>
- *Kost, C., Foss, P., & Lenzini, 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>
- Mahdi, H. S., Mohsen, M. A., & Almanea, M. (2024). Multimedia glosses and second language vocabulary learning: A second-round meta-analysis. *Acta Psychologica*, 248, 104341. <https://doi.org/10.1016/j.system.2024.103273>
- Mayer, R. (2001). *Multimedia learning*. Cambridge University Press.
- Mayer, R. (2009). *Multimedia learning (2nd ed.)*. Cambridge University Press.
- Mohsen, M., & Balakumar, M. (2011). A review of multimedia glosses and their effects on L2 vocabulary acquisition in CALL literature. *ReCALL*, 23(2), 135–159. <https://doi.org/10.1017/S095834401100005X>
- Mohsen, M. A., Mahdi, H. S., & Alkhamash, R. (2024). Multimedia glosses and their impact on second language vocabulary acquisition: Insights from a meta-analysis and document co-citation analysis. *Innovation in Language Learning and Teaching*, 18(2), 109–124. <https://doi.org/10.1080/17501229.2023.2236084>
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
- Nation, I. S. P. (2013). *Learning vocabulary in another language (2nd edition)*. Cambridge University Press.

- Nation, I., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning.
- Norris, J., & Ortega, L. (2006). (Eds.), *Synthesizing research on language learning and teaching*. John Benjamins Pub.
- Plonsky, L. & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912. <https://doi.org/10.1111/lang.12079>
- Plonsky, L. & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky. (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge.
- *Ramezanali, N., & Faez, F. (2019). Vocabulary learning and retention through multimedia glossing. *Language Learning & Technology*, 23(2), 105–124. <https://doi.org/10.125/44685>
- 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>
- Rassaei, E. (2018). Computer-mediated textual and audio glosses, perceptual style and L2 vocabulary learning. *Language Teaching Research*, 22(6), 657–675. <https://doi.org/10.1177/1362168817690183>
- Rassaei, E., & Folse, K. (2024). Effects of L1 and L2 word-level vs. L2 sentence-level glosses on vocabulary learning. *System*, 122, 103273. <https://doi.org/10.1016/j.system.2024.103273>
- Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effect models. In H. Cooper, L. Hedges, & J. Valentine. (Eds.), *The handbook of research synthesis and meta-analysis (2nd ed.)* (pp. 295–316). Russell Sage Foundation.
- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.
- Sadeghi, K., Khezrlou, S., & Modirkhameneh, S. (2017). Calling Iranian learners of L2 English: Effect of gloss type on lexical retention and reading performance under different learning conditions. *Journal of Research in Reading*, 40(S1). <https://doi.org/10.1111/1467-9817.12088>
- Schmidt, R. (2001). Attention. In P. Robinson. (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.
- Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. Hedges, & J. Valentine. (Eds.), *The handbook of research synthesis and meta-analysis (2nd ed.)* (pp. 257–278). Russell Sage Foundation.
- Taylor, A. (2009). CALL-based versus paper-based glosses: Is there a difference in reading comprehension? *CALICO Journal*, 27, 147–160. <http://doi.org/10.11139/cj.27.1.147-160>
- Taylor, A. (2013). CALL versus paper: In which context are L1 glosses more effective? *CALICO Journal*, 30(1), 63–81. <https://doi.org/10.11139/cj.30.1.63-81>
- *Teng, M. F. (2023). The effectiveness of multimedia input on vocabulary learning and retention. *Innovation in language learning and teaching*, 17(3), 738–754. <https://doi.org/10.1080/17501229.2022.2131791>
- Teng, M. F. & Reynolds, B. (2024). (Eds.), *Researching incidental vocabulary learning in a second language*. Routledge.
- *Teng, M. F., & Zhang, D. (2023). The associations between working memory and the effects of multimedia input on L2 vocabulary learning. *International Review of Applied Linguistics in Language Teaching*, 61(3), 1021–1049. <https://doi.org/10.1515/iral-2021-0130>
- *Teng, M. F., & Zhang, D. (2024). Vocabulary learning in a foreign language: multimedia input, sentence-writing task, and their combination. *Applied Linguistics Review*, 15(5), 2123–2148. <https://doi.org/10.1515/applirev-2022-0160>
- Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 69, 559–599. <https://doi.org/10.1111/lang.12343>
- Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. Hedges, & J. Valentine. (Eds.), *The handbook of research synthesis and meta-analysis (2nd ed.)* (pp. 129–146). Russell Sage Foundation.
- *Wang, S., Liu, T., & Lee, C. I. (2023). The efficacy of CALL glossing on EFL learners’ idiom learning. *Acta Psychologica*, 241, 104060. <https://doi.org/10.1016/j.actpsy.2023.104060>

- 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>
- Webb, S., & Nation, I. (2017). *How vocabulary is learned*. Oxford University Press.
- 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>
- *Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, 13(2), 48–67. <http://dx.doi.org/10125/44180>
- *Yeh, Y., & Wang, C. (2003). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journal*, 21(1), 131–144. <https://doi.org/10.1558/cj.v21i1.131-144>
- *Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, 10, 85–101. <http://dx.doi.org/10125/44076>
- *Yoshii, M., & Flaitz, J. (2002). Second language incidental vocabulary retention: The effect of text and picture annotation types. *CALICO Journal*, 20(1), 33–58. <https://doi.org/10.1558/cj.v20i1.33-58>
- Yun J. W. (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>
- Zhang, C., & Ma, R. (2024). The effect of textual glosses on L2 vocabulary acquisition: A meta-analysis. *Language Teaching Research*, 28(3), 967–986. <https://doi.org/10.1177/13621688211011511>
- *Zou, D., & Teng, M. F. (2023). Effects of tasks and multimedia annotations on vocabulary learning. *System*, 115, 103050. <https://doi.org/10.1016/j.system.2023.103050>

Author Bio

Yang Qiwei is a Ph. D. candidate from School of Foreign Languages and Literature, Beijing Normal University. His research interests include foreign language vocabulary learning and teaching, second language acquisition and language testing. He is currently doing his doctoral program in vocabulary learning from L2 reading.

Appendix A

Some excluded studies and the reasons

Author/Year	Title	Reasons
Rassaei, 2018	Computer-mediated textual and audio glosses, perceptual style and L2 vocabulary learning	This study did not involve any pictures for gloss.
Abraham, 2007	Second language reading comprehension and vocabulary learning with multimedia	The focus of the comparison is not put on the different modes.
Ko, 2005	Glosses, comprehension and strategy use	The dependable variable is the reading comprehension rather than the vocabulary recall or recognition.
Plass, Chun, Mayer, & Leutner, 2003	Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities	There is not adequate information for calculation of the effect size.

Appendix B

Coding of each study

No.	Author	Year	Appeared in	Age group	Age	Group size	L1	L2	L2 proficiency	Text Type
1	Acha	2009	Journal of Educational Technology	Elementary school	8.5	135	Spanish	English	beginner	Story
2	Akbulut	2007	Instructional science	University	None	69	Turkish	English	advanced	Playing with fire
3	Al-Seghayer	2001	Language Learning and Technology	University	None	30	Arabic, Korean	English	intermedia	Narrative
4	Brimshaw & Smith	2017	Computer assisted language learning	University	None	33	Not reported	English	high-intermedia	Story
5	Warren, He, De	2017	System	University	19-20/21-24/19-20	125	Chinese/Malaysian	English	eight years	Narrative
6	Chun Plass	1996	Modern language Journal	University	20.67/21.9/21.05	160	English	German	second year	Story
7	Hey Ramonda	2012	Language teaching research	University	None	87	Not reported	Spanish	first year	
8	Jones Plass	2002	Modern language Journal	University	None	171	English	French	beginner	Story
9	Kim Gilman	2008	Instructional technology and systems	Middle school	14	172	Korean	English	Not reported	
10	Cost Foss Lenzi	1999	Foreign language annals	University	None	56	English	German	first year	Story
11	Yanguas	2009	Language learning and technology	University	None	94	English	Spanish	second year	Expository
12	Yeh Wang	2003	CALICO	University	None	82	Chinese	English	six year	expository
13	Yoshii	2006	Language Learning and Technology	University	None	195	Japanese	English	Not reported	Story
14	Yoshii Flaitz	2002	CALICO	University	24.6	151	8 languages	English	beginner+intermedia	Story
15	Özkemak & Erçetin	2018	ReCALL	University	None	88	Turkish	English	elementary	Story
16	Amazanali & Farman	2019	Language Learning and Technology	Mixed	17.56	132	Farsi	English	intermedia	Story
17	Sadeghi et al	2016	Journal of research in reading	University	18-21	135	Farsi	English	Not reported	Story
18	Teng	2023	Journal of language learning and technology	University	18-20	125	Chinese	English	Not reported	Expository
19	Teng & Zhang	2023	Journal of review of applied linguistics	University	18-20	95	Chinese	English	Not reported	Expository
20	Teng & Zhang	2024	Applied linguistics review	University	18-20	235	Chinese	English	Not reported	Expository
21	Zou & Teng	2023	System	University	25-35	360	Chinese	English	intermedia	Story
22	Wang, Liu, Lee	2023	Acta psychologica	University	18-21	99	Chinese	English	intermedia	Story

Appendix C

Effect size and confidence interval of each study

