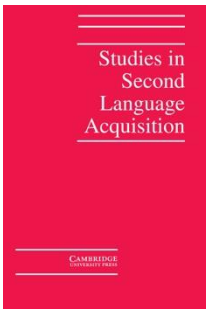
Accepted for publication in *SSLA* (Cambridge University Press)



How do Different Forms of Glossing Contribute to L2 Vocabulary Learning from Reading? A Meta-Regression Analysis

Akifumi Yanagisawa, Stuart Webb, Takumi Uchihara
(University of Western Ontario)

Abstract

This meta-analysis investigated the overall effects of glossing on L2 vocabulary learning from reading and the influence of potential moderator variables: gloss format (type, language, mode) and text and learner characteristics. A total of 359 effect sizes from 42 studies (N = 3802) meeting the inclusion criteria were meta-analyzed. The results indicated that glossed reading led to significantly greater learning of words (45.3% and 33.4% on immediate and delayed posttests, respectively) than non-glossed reading (26.6% and 19.8%). Multiple-choice glosses were the most effective, and in-text glosses and glossaries were the least effective gloss types. L1 glosses yielded greater learning than L2 glosses. We found no interaction between language (L1, L2) and proficiency (beginner, intermediate, advanced), and no significant difference among modes of glossing (textual, pictorial, auditory). Learning gains were moderated by test formats (recall, recognition, other), comprehension of text, and proficiency.

Keywords: Glossing, Annotation, Incidental vocabulary learning, Reading, Meta-analysis

How do Different Forms of Glossing Contribute to L2 Vocabulary Learning from Reading? A Meta-Regression Analysis

Research has demonstrated that L2 students can incidentally learn unknown words from reading (e.g., Day, Omura, & Hiramatsu, 1992; Horst, Cobb, & Meara, 1998; Pigada & Schmitt, 2006; Webb & Chang, 2015). However, studies tend to show that learning gains from reading are relatively small. This is mainly because inferring the meanings of unknown words in context is challenging (e.g., Nassaji, 2003). For example, to successfully guess a word's meaning, learners need to know the majority of the words in a text to understand the context (Hu & Nation, 2000; Laufer, 1989; Schmitt, Jiang, & Grabe, 2011), and in turn provide the greatest potential of inferring the unknown word (Liu & Nation, 1985). One solution to this issue is to provide glosses for unknown words (Webb & Nation, 2008). Research has demonstrated that glossing or annotating unfamiliar words not only promotes L2 reading comprehension but also increases learning of the glossed words (e.g., Hulstijn, Hollander, & Greidanus, 1996; Jacobs, Dufon, & Hong, 1994; Watanabe, 1997).

Many studies have compared different forms of glossing to determine how glosses can maximize vocabulary learning from reading. Some studies have compared glossing types such as marginal glosses, multiple-choice glosses, in-text glosses, and hyperlinked glosses (Nagata, 1999; Watanabe, 1997). Other studies have examined which language (L1 or L2) should be used to provide the meanings in glosses (Jacobs et al., 1994; Ko, 2012; Yoshii, 2006), or the mode (textual, visual, auditory) in which glosses are presented (Kost, Foss, & Lenzini, 1999; Yoshii & Flaitz, 2002). However, despite a great deal of research on the efficacy of different forms of glossing, there tend to be contradictory conclusions, and there remains a lack of consensus on the relative values of the different approaches to glossing.

One way to deepen our understanding of how glosses should be provided is to conduct a meta-analysis. Individual studies are restricted, for example, by their research foci, participant population, target languages, and materials. In contrast, meta-analysis allows the examination of how the effects of glossing vary in relation to variables such as gloss formats (gloss type, language, mode), characteristics of texts (e.g., learner targeted texts, proportion of glossed words) and learners (e.g., institutional level, proficiency) by synthesizing studies conducted in different contexts. Results from meta-analyses are usually more reliable because they are based on the results of multiple studies. Furthermore, the increased sample sizes through meta-analysis provide a higher statistical power (Hunter & Schmidt, 2004), which enables the re-examination of the value of variables that have been rejected due to low statistical power. The present meta-analysis aims to investigate (1) the overall learning gain from glossed reading, and (2) how glossing effects vary based on format (i.e., type, language, mode) and the characteristics of texts and learners. The current meta-analysis may provide substantial pedagogical implications as to how glosses should be provided for language teachers and material designers.

Background

Generally, glossing in the L2 learning context refers to providing L1 translations, L2 synonyms, or short explanations of unfamiliar words in a text (Bowles, 2004; Nation, 2013). Many studies have demonstrated that glossed reading leads to greater vocabulary learning than non-glossed reading (e.g., Hulstijn et al., 1996; Jacobs et al., 1994; Ko, 2012; Watanabe, 1997).

Glossing has many attractive features. For example, glosses are easier to access than dictionaries, and their presence in a text allows learners to continue on with the reading process with little interruption (Yanguas, 2009). This may lead learners to check the meanings of words

more frequently using glosses than other resources (Hulstijn et al., 1996). Also, glosses enhance learners' noticing of unknown target language items (Bowles, 2004; Rott, 2005; Yanguas, 2009), and ensure appropriate form-meaning connections (Nation, 2013). This may help readers to learn the meanings of target words efficiently. Furthermore, glossed reading provides opportunities for learners to foster their autonomy by lessening the need to depend on teachers for explanations (Jacobs et al., 1994; Nation, 2013). Researchers' focus has therefore shifted from whether glosses promote vocabulary learning to how glosses should be provided to maximize their effectiveness (Azari, 2012; Mohsen & Balakumar, 2011; Yoshii, 2006).

Gloss Type

There are many types of glosses, which can be roughly categorized into non-interactive glosses and interactive glosses. Non-interactive glosses are inserted at a specific place (e.g., margin of a text) and are clearly presented for learners' use. Non-interactive glosses include marginal (provided in the margin of a text, usually at the right hand or bottom), interlinear (provided between the lines of a text), in-text glosses (provided next to the word in a text), and glossaries (provided at the end of the text or as a separate paper in the form of a list). Paper reading materials are still frequently used, and these non-interactive glosses are adopted as common approaches among language teachers.

Interactive glosses require learners to take action. One example of interactive glosses is hyperlinked glosses, which require learners to click on or put a mouse cursor over a word to see the gloss. Webb and Nation (2017) argue that hyperlinked glosses are particularly useful because they allow for retrieval of the word meaning. While non-interactive glosses tend to be provided in places that can be easily seen, hyperlinked glosses allow learners to retrieve the meanings of glossed words before opening the glosses. Another example of an interactive gloss type is multiple-choice glosses. Multiple-choice glosses tend to be provided at the margin of texts. Instead of simply providing the corresponding meanings of words, there are several options for a target word meaning. Learners are required to read texts carefully and select the appropriate option that fits the context. Hulstijn (1992) argues that multiple-choice glosses cause deeper processing of glossed words, thus leading to greater vocabulary learning.

Earlier studies have compared the relative effectiveness of different types of glosses. Multiple-choice glosses are one of the most frequently investigated gloss types. Studies comparing multiple-choice glosses and single-translation glosses have reported mixed results. While some studies (Hulstijn, 1992; Miyasako, 2002; Watanabe, 1997) did not find clear differences between the two types of glosses, others (Nagata, 1999; Rott, 2005) found that multiple-choice glosses led to greater vocabulary learning than single-translation glosses. Furthermore, Yoshii (2013) found that single-translation glosses led to greater learning than multiple-choice glosses.

In-text glosses are thought to be the least effective type of glossing (Nation, 2013; Schmitt, 2008). Watanabe (1997) did not find a clear difference between an in-text gloss condition and a non-glossed condition. However, Cheng and Good (2009) showed that an in-text gloss condition yielded greater learning than a non-glossed condition. For this reason, it might be too early to conclude that in-text glosses have no effect on learning.

Gloss Language

Whether the meanings of words should be glossed in the L1 (translations) or L2 (definitions or synonyms) has been a long-standing question in vocabulary learning. One

advantage of L1 glosses relates to providing short and clear meanings for unknown words regardless of learners' proficiency levels. Learners are less likely to misinterpret the meanings of glossed words when L1 glosses are presented compared to when words are glossed with L2 definitions. However, L2 glosses also have advantages. L2 glosses are more easily provided when learners who have different L1 backgrounds. Furthermore, L2 glosses increase the amount of input in the target language. One shortcoming of L2 glosses, however, requires teachers to ensure that every word in the gloss is clearly comprehensible for all learners. This increases the burden on teachers. Failing to control the comprehensibility of glosses cancels out the benefit of glossing.

Earlier studies comparing the effects of L1 and L2 glosses (Jacobs et al., 1994; Ko, 2012; Miyasako, 2002; Yoshii, 2006) have presented inconsistent findings. Jacobs et al. (1994), Ko (2012), and Yoshii (2006) found no clear difference between L1 and L2 glosses. In contrast, Fang (2009), Xu (2010), and So (2010) found an advantage of L1 glosses over L2 glosses. Furthermore, different gloss languages might affect retention of words differently. Yoshii (2006) found that scores on a meaning recall test showed different forgetting rates for words glossed in the L1 and L2 and indicated that the vocabulary gains in the L1 gloss condition declined more sharply over time than the L2 gloss condition.

Combining both the L1 and L2 in glosses is another approach. Some studies have investigated the effect of L1 plus L2 glosses in comparison to L1 and L2 glosses on their own (Azari, Abdullah, Heng, & Hoon, 2012; Ko, 2017; Salehi & Naserieh, 2013; Xu, 2010). The results of earlier studies are again inconsistent. Ko (2017) and Azari et al. (2012) found that L1 plus L2 glosses were superior to either L1 or L2 glosses alone. Xu (2010) found that L1 glosses led to more learning than L1 plus L2 glosses on an immediate posttest, but on a delayed posttest, the L1 plus L2 glosses contributed to larger gains than L1 glosses. In contrast, Salehi and Naserieh (2013) did not find any advantage of L1 plus L2 glosses over L1 glosses or over L2 glosses. However, all of these conditions yielded greater learning than a non-glossed condition.

Many researchers argue that learners' L2 proficiency may moderate the effectiveness of the language used for glossing (Ko, 2012, 2017; Yoshii, 2006). For example, Nation (2013, p. 246) suggests that L1 glosses should be provided for beginners, while either L1 or L2 would work for advanced learners. Ko (2017) tested the effectiveness of three different conditions with a control condition (i.e., L1 gloss, L2 gloss, and L1 plus L2 gloss, no-gloss) with Korean EFL learners. The findings indicated that the L1 and L1 plus L2 glosses were effective for lower proficiency learners, while the L2 and L1 plus L2 glosses were effective for higher proficiency learners. However, because few individual studies directly examined the interaction between gloss languages and students' proficiency levels, it is difficult to reach a conclusion just by looking at the results of individual studies. Alternatively, meta-analysis may allow us to investigate the effects of L2 proficiency level on the language of the gloss, by looking at the results of multiple studies that have recruited participants with different proficiency levels.

Gloss Modalities

Glosses can be presented in different modalities. Widely-used modalities include text, picture, video, auditory, and the combination of one or more of these modalities. The majority of the studies demonstrated that pictorial and textual glosses were equally effective (Kost et al., 1999; Yanguas, 2009; Yoshii & Flaitz, 2002). Rassaei (2017) compared textual glosses and audio glosses. The results showed that audio glosses yielded greater learning than textual glosses both on the productive recall test and recognition test regardless of test timings.

Studies have also focused on whether combining different modalities (usually textual and pictorial) enhances the glossing effect (Boers, Warren, He, & Deconinck, 2017; Chun & Plass, 1996; Warren, Boers, Grimshaw, & Siyanova-Chanturia, 2018; Yeh, Wang, & Tsing, 2003). Yun's (2011) and Vahedi et al.'s (2016) earlier meta-analyses on the effect of glossing on vocabulary focused on the effects of multimodal glosses compared to single-mode glosses. However, the comparison between single-mode glosses and multimodal glosses does not indicate which specific gloss mode is more advantageous for vocabulary learning. Meta-analyzing earlier studies examing effects of single-mode glosses may shed light on the relative effectiveness of different types of gloss modes.

Rationale of the Present Study

Due to the inconsistency of results in the literature, there is little consensus on how glosses should be provided. One approach to disentangling this incongruity is to conduct a meta-analytic review. Several attempts to synthesize studies investigating the effects of glossing on vocabulary learning from reading have already been made in earlier meta-analyses (Abraham, 2008; Vahedi et al., 2016; Yun, 2011; see also Taylor, 2006, 2009, 2013, and 2014 for effects of glossing on reading comprehension). Abraham (2008) synthesized 11 studies to investigate the overall effects of CALL (computer-assisted language learning) glosses on vocabulary learning and reading comprehension. Abraham examined the influence of moderator variables: proficiency, text type, and test. The study did not investigate how different gloss types, language, and modes affected vocabulary learning. Additionally, as Abraham notes, the meta-analysis results were based on a small number of studies (only six studies were analyzed as to the glossing effect on vocabulary learning), weakening the statistical power to test moderator variables and the reliability of its results.

Vahedi et al. (2016) and Yun (2011) investigated the effect of multimodal (textual plus pictorial) glosses compared to single-mode (textual) glosses. Both of the studies compared treatment groups accessing multiple hypertext glosses with control groups accessing single-mode glosses in vocabulary learning through reading and found that multimodal glosses were found to be moderately more effective than single-mode glosses (Hedges' g = 0.46 in Vahedi et al; and Hedges' g = 0.84 in Yun). These studies indicated that combining different gloss modes are more effective than single-mode glosses; however, we still do not know the extent to which single mode glosses promote learning in its own right. This is surprising since it seems important to first understand the strengths and weaknesses of single-mode glosses before exploring the complex effects of the combinations of different gloss modes (i.e., multimodal glosses). Therefore, in our meta-analysis, we focused on various formats of single-mode glosses.

All three earlier meta-analyses on glossing (i.e., Abraham, 2008; Vahedi et al. 2016; Yun, 2011) noted that their results may suffer from weak statistical power and reliability due to the small number of studies included. In order to increase the number of studies included, we used relative learning gain (proportion of unknown words learned) as standardized effect sizes (ESs) by following Swanborn and de Glopper's (1999) meta-analysis of incidental L1 vocabulary learning from reading. This allows more studies to be included in the analysis because we can include not only studies that employed true control groups (i.e., participants who read texts without glosses) but also studies that compared different types of glosses (e.g., studies comparing L1 glosses and L2 glosses) without having true control groups. Using a relative learning gain also provides more comprehensible ESs (Swanborn & de Glopper, 1999) representing the proportions of unknown words learned by participants with and without the provision of

glossing. Furthermore, to account for variances between studies, this meta-analysis adopted three-level meta-regression (de Vos, Schriefers, Nivard, & Lemhöfer, 2018; Lee, Warschauer, & Lee, 2018; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2015). This allows the examination of differences between different forms of glossing used within each individual study as well as differences of different forms of glossing aggregated across all included studies. This analysis provides a more reliable estimation of the relative effectiveness between different glossing formats (e.g., Konstantopoulos, 2011).

Meta-analysis also allows the examination of variables that were not the focus of each individual study. By conducting moderator analyses, we examined the relationships between learning gains and the features of studies, such as the characteristics of reading materials and participants. Investigating these variables may offer a more precise and transparent picture of how glosses affect learning in different contexts. Additionally, by including test format as a variable, this study may provide a more accurate assessment of the extent to which vocabulary knowledge is gained through glossed reading.

The current study was guided by the following five research questions.

1. What is the overall effect of glossed reading on L2 incidental vocabulary learning from reading?
2. What are the relative effects of gloss type?
3. What are the relative effects of gloss language?
4. What are the relative effects of gloss mode?
5. To what extent do text and learner characteristics moderate glossing effects?

Method

Literature Search

We used several databases in order to search for studies to include in the meta-analysis—Education Resources Information Centre (ERIC), PsycINFO, Linguistics and Language Behavior Abstracts (LLBA), ProQuest Global Dissertations, and Google Scholar—by using various combinations of the following keywords: *gloss, glosses, glossing, vocabulary, reading, annotation, annotate, annotating, L2, foreign, and second.* This process identified 1001 reports published until and including August 2018, all of which were retrieved and carefully reviewed.

To ensure the comprehensiveness of the literature search, we also inspected the reference sections of the previous meta-analyses (Abraham, 2008; Vahedi et al., 2016; Yun, 2011), review articles on glossing effects on vocabulary learning (Azari, 2012; Mohsen & Balakumar, 2011), and studies identified through the electronic database search. Studies with relevant topics were also retrieved.

Inclusion and Exclusion Criteria

The following seven criteria were employed to determine which studies to include in the analysis.

1. Experimental or quasi-experimental studies investigating glossing effects on L2 vocabulary learning by comparing reading with glossing groups and a reading without glossing group, or by comparing different types of single-mode glossing were included.
2. Studies in which participants read a text or multiple texts were included. Studies focusing on sentence reading were not included.

3. Studies that used a between-participants design were included. Studies that used a within-participants design (e.g., Al-Seghayer, 2001) were excluded due to the limited number of studies using a within-participants design. This decision was based on the suggestion that blending two different research designs in a meta-analysis tends to produce biased results (Plonsky & Oswald, 2015).
4. Studies that only used unknown words as target words or controlled participants' prior knowledge of the target words either by using nonsense words or pretest scores were included. We excluded studies reporting that participants already knew a considerable number of the target words before the treatment by administering pretests, but did not report the pretest scores (e.g., Jalali & Neiriz, 2012).
5. Studies that focused on incidental vocabulary learning (participants were not forewarned about vocabulary posttests, Hulstijn, 2001) were included. Studies in which participants were told that they would be tested on vocabulary or explicitly told to learn vocabulary (e.g., Holley & King, 1971), as well as studies where participants were repeatedly tested before and after the reading sessions (e.g., AbuSeileek, 2013) were excluded. As such, we aimed to exclude studies in which there was the possibility that participants focused on remembering glossed target words instead of reading texts for comprehension (Hulstijn, 1992). One study in which participants engaged in word-focused activities (i.e., So, 2010) was also excluded.
6. Studies reporting enough statistical information (i.e., mean, *SD*, number of participants tested) required for meta-analysis were included.
7. Articles written in English were included.
8. Studies were excluded when different target words were used in the different learning conditions, or there was insufficient information to interpret the methodologies (e.g., not reporting how vocabulary knowledge was measured).
9. Studies that focused on the effect of corpus-based glosses (e.g., Lee, Warschauer, & Lee, 2017) were excluded. This is because corpus-based glosses provide corpus-extracted sentences for glossed words and require readers to infer the meanings of target words, which is quite different from other gloss types where form-meaning connections are provided.

Given the inclusion of unpublished studies (e.g., M.A. and Ph.D. theses, conference presentations), securing the quality of the included studies is vital. We strictly excluded studies that failed to clearly describe how learning gains were measured or how participants' responses on tests were scored. Forty-two studies (N = 3802) providing 359 posttest scores satisfied all of the criteria (see On-line Supplementary Material Appendix 2 for the flowchart for inclusion and exclusion of the studies following the PRISMA 2009 flow diagram, Moher, Liberati, Tetzlaff, Altman, & Group, 2009). These 42 studies comprised 31 journal articles, three conference presentations, three bulletin reports, three Ph.D. theses, one M.A. thesis, and one online report (see On-line Supplementary Material Appendix 3 for basic information about the included studies).

Coding

All 42 studies meeting the selection criteria were coded for several variables: outcome variables, glossing variables, textual and learner characteristics, as well as study identifier (e.g., author, year, experiment number, participant group number).[1]

Outcome variables. For dependent variables, mean posttest scores, *SD*s, number of participants tested, and test timing (immediate or delayed) were coded. To account for the strength of vocabulary knowledge measured (Yanagisawa & Webb, 2019), we followed earlier meta-analyses on L2 vocabulary learning (Uchihara, Webb, & Yanagisawa, 2019; de Vos et al., 2018) and coded test format as either: (a) recall (b) recognition, or (c) other tests. de Vos et al.'s (2018) meta-analysis of L2 vocabulary learning from spoken input grouped vocabulary tests into two types, recognition and recall, based on its sensitivity. Recognition tests included form recognition (knowledge of word forms, i.e., spellings), and receptive and productive recognition (knowledge of form-meaning links). Recall tests included receptive and productive recall (knowledge of form-meaning links). We added one more category, *other tests*, where participants' vocabulary knowledge was tested beyond the form-meaning connection and involved measurement of participants' use of target words. Fill-in-the-blank tests and the Vocabulary Knowledge Scale (VKS) were coded as other test formats.

Glossing variables. Glossing was coded for type, language, and modality. Gloss type was coded as either: marginal, multiple-choice, glossary, in-text, interlinear, hyperlinked, or no-gloss. There were no other types of glosses identified among the included studies. Gloss languages were coded as L1, L2, or L1 plus L2. Gloss modalities were coded as textual, pictorial, or auditory. Other gloss languages or modalities were not identified in the included studies.

Text characteristics. Five variables were coded for text characteristics: L2 learner targeted material, comprehensibility, target word percentage, text type, and CALL use. For L2 learner targeted material, we coded texts as either targeted for L2 learners or native speakers. Reading materials were coded as L2 learner targeted material when they were written or edited by researchers and when researchers used texts of textbooks or workbooks written for L2 learners. When the title of the book (e.g., textbooks and workbooks) was reported, we searched on the Internet to determine whether the book was written for L2 learners or native speakers. When it was not clear whether the material was written for L2 learners or native speakers, the study was excluded from this analysis.

Some studies investigated participants' reading comprehension using recall tests or multiple-choice questions. We coded and standardized the test scores (dividing the scores by the maximum score of each test then multiplying the resultant scores by 100), which indicates how well participants understood the reading material or the difficulty of the reading material.

Target word percentage was calculated by dividing the total number of target words by the total number of words participants read then multiplying by 100. Target word percentage can roughly be interpreted as the percentage of words that the researchers believed the learners did not know in texts.

Text type was coded as either narrative or expository. CALL use was coded as either yes (i.e., reading material was presented on a computer screen), or no (i.e., on paper).

Learner characteristics. Learners' L2 proficiency levels and institutional levels were coded. For learners' proficiency, we coded them as beginner, intermediate, or advanced. Institutional level was coded as secondary school, university, or language school. No other institution was identified in the included studies.

Coding procedure. Following earlier meta-analyses and suggestions (Plonsky & Oswald, 2015), two of the authors of this study engaged in the coding process to enhance the reliability of the coding. First, the two authors coded five studies separately using the coding scheme. The coding agreement between the two coders was calculated by using Cohen's kappa and showed high agreement rate at $\kappa = .97$. All discrepancies were discussed until achieving consensus. Finally, one of the authors carefully coded the rest of the studies, while the other author randomly coded 28.6% (12 out of 42) of the remaining studies. This time, there were no discrepancies between coders.

Effect Size Calculation

Following Swanborn and de Glopper's (1999) meta-analysis on vocabulary learning, we adopted the proportion of unknown words learned as ESs using the following formula (see also, Horst et al., 1998; Webb & Chang, 2015):

$$\text{ES} = \frac{\text{Mean posttest score} - \text{Mean pretest score}}{\text{Maximum posttest score} - \text{Mean pretest score}}$$

When studies used target words that were all unknown to participants, mean pretest score was set as 0. When studies used control group (i.e., a group that only took the same vocabulary test without any exposure to target words) to account for participants prior knowledge of target words, the control group's mean posttest score was used as the mean pretest score (see On-line Supplementary Material Appendix 4 for the detailed calculation formulas).

In order to appropriately weight ESs, we calculated the sampling variance using the reported *SD*s of posttest scores that were converted into proportions. The formula in Hox (2010, p. 209) was used, where *s* refers to a proportion-translated *SD* of posttest scores:

$$\text{Sampling variance} = s^2/n$$

In Swanborn and de Glopper (1999), only number of participants was used to calculate sampling variance while ignoring the test score variance (which is reported as *SD* or *SE* in individual studies). In order to obtain more precise information about the sampling variance to enhance the accuracy of the analysis, the current meta-analysis used reported *SD* of posttest scores to calculate sampling variance. The detailed analytic approach regarding ES and sampling variance calculation, as well as calculation formulas for ESs and sampling variances, can be found in the On-line Supplementary Material Appendix 4.

Overall, the included 42 studies reported 359 posttest scores, all of which were transformed into ESs and included in the analysis.

Analysis

As described previously, we converted all reported posttest scores into a relative learning gain (i.e., proportion of the unknown words learned) and used this as ESs. A relative gain is a widely used index to standardize the learning gain among the vocabulary research (e.g., Horst et al., 1998; Swanborn & de Glopper, 1999; Webb & Chang, 2015). Using relative learning gains allows us to compare learning gains from different conditions of different studies on the same scale. Furthermore, we used a three-level meta-analysis (e.g., Cheung, 2014; de Vos et al., 2018; Lee et al., 2018) to model three different sources of variances with sampling variance being the first level, variance between the different ESs from the same study being the second, and variance between studies being the third. The three-level meta-analysis is a technique that considers dependencies of ESs (i.e., which ESs are produced by the same studies) so that variances within studies as well as variances between studies are examined at the same time.

This technique produces more reliable and robust estimations when examining the different glossing conditions compared within each study (e.g., Konstantopoulos, 2011). Furthermore, quite a few studies reported multiple ESs that are dependent due to sampling error variance (i.e., the same participants were tested with different measurements or repeatedly tested at different times). In order to appropriately deal with the potential Type I error inflation due to these dependent ESs, we applied the cluster-robust variance estimation (Hedges, Tipton, & Johnson, 2010) with small-sample adjustments (Tipton, 2015; Tipton & Pustejovsky, 2015). All the analyses were conducted in the R statistical environment (R Core Team, 2017) using the metafor package (Viechtbauer, 2010) and the clubSandwich package (Pustejovsky, 2018). An *F* test (i.e., Wald-test) was conducted when there was more than one level in a predictor variable included in a statistical model. ESs of immediate posttests and delayed posttests were analyzed separately. We set the significance level at 5%. *P*-values lower than 0.10 were also reported as of marginal significance. This did not mean to imply that the null hypothesis was rejected, but to report the trend of the data.

Analysis procedure. The ESs of immediate posttests and delayed posttests were analyzed separately. For Research Question 1, to investigate the overall effect of glossed reading, meta-regression models with a Gloss variable (gloss vs. non-gloss) that predicts the ESs with and without intercept were fitted with the data. Then, to investigate whether the glossing effect varied based on test formats, the interaction term between the Gloss variable and Test format was added to the model.

For Research Questions 2, 3, and 4 comparing different formats of glosses (type, language, mode) to investigate whether different formats led to significantly greater learning than non-gloss reading, meta-regression models with a Gloss format variable (type, language, mode) that predicts ESs were administered. Test format was also inserted as a covariate to control for the potential bias due to the test format (e.g., less effective gloss types might not have been measured with more demanding tests). Non-glossed condition and recall test format were set as reference levels. Next, to investigate the relative effectiveness between different formats, multiple comparisons were administered by changing the reference levels (de Vos et al., 2018).

For Research Question 5 investigating the effects of text and learner characteristics, to determine whether each variable correlated with ES, each predictor variable was inserted into a meta-regression model with a Gloss variable (gloss vs. non-gloss) and Test format. Next, to investigate whether the effect of each variable changed based on having glosses or not, an interaction term between each variable and Gloss variable was added to the first meta-regression model. In order to avoid biases caused by outliers, for analyzing numerical variables (i.e., comprehensibility and target word percentage), ESs for each variable deviated 2 *SD* from the mean were excluded.

Sensitivity Analyses. Potential publication biases and outliers of ESs influencing the overall ESs were checked using Egger's tests (Egger, Smith, Schneider, & Minder, 1997), funnel plots, and Cook's distance (de Vos et al., 2018; Viechtbauer & Cheung, 2010). Examining funnel plots showed that an ES based on Zhao and Ren's (2017) immediate posttest score of the non-glossed reading condition, the low-proficiency group indicated a very small variance (*SD* = 0.08, maximum test score = 24) that greatly deviated from other ESs, so this participant group was excluded from the analysis. No other obvious publication biases or outliers were identified. We also analyzed whether learning gains or glossing effect influenced publication status. Those

preliminary analyses did not find any publication biases. Furthermore, other methodological variables (e.g., participant allocation method, delayed posttest timing) were also examined. The detailed results of these diagnostic tests, as well as methodological variables, can be found in the On-line Supplementary Material Appendix 5.

Results

Research Question 1: What is the Overall Effect of Glossed Reading on L2 Incidental Vocabulary Learning from Reading?

A total of 30 studies included non-gloss reading conditions with 89 posttest scores, and 42 studies included glossed reading conditions with 259 posttest scores. Table 1 shows the overall effect of learning gains separately for glossed reading and non-glossed reading. For immediate posttests, glossed reading led to 45.3% (95% CI [38.7, 51.8]) of the unknown words were learned (b = 0.453, SE = 0.032, p < .001). Non-glossed reading yielded the learning of 26.6% (95% CI [20.5, 32.7]) of the unknown words (b = 0.266, SE = 0.030, p < .001), and this was significantly lower than glossed reading (b = -0.187, SE = 0.022, p < .001).

For delayed posttest scores, glossed reading led to the learning of 33.4% (95% CI [27.0, 39.8]) of the unknown words (b = 0.334, SE = 0.032, p < .001). Non-glossed reading led to the learning of 19.8% (95% CI [13.1, 26.5]) of unknown words (b = 0.198, SE = 0.033, p < .001). The difference between glossed reading and non-glossed reading was significant (b = -0.136, SE = 0.021, p < .001), showing that the glossing effect was retained through to the delayed posttests.

Table 1
The Overall Effect of Glossed-reading

	Immediate					Delayed				
	k	*n*	Mean ES (%)	CI	*p*	*k*	*n*	Mean ES (%)	CI	*p*
Gloss	39	154	45.3	38.7, 51.8	< .001	36	113	33.4	27.0, 39.8	< .001
Non-gloss	27	45	26.6	20.5, 32.7	< .001	28	47	19.8	13.1, 26.5	< .001

Note. *k* = number of studies, *n* = number of effect sizes, CI = 95% confidence interval, Mean ES = weighted mean effect sizes converted into a percentage from proportion for the sake of interpretability. *p* = *p*-value for significant test. Pseudo-R^2 (Raudenbush, 2009) showed that 11% and 9.4 % of the variance in the ESs was explained for the immediate and the delayed posttests, respectively.

How do learning gains differ in relation to test format? Learning gain from reading may vary significantly across the different types of vocabulary measurements used in each study. First, to determine whether vocabulary learning from reading, in general, differs in the format of measurements, we entered the main effects of Gloss and Test format into a meta-regression model (see Table 2 for the mean effect sizes for each test format). Second, to determine whether glossing promoted the learning of a specific aspect of vocabulary knowledge, the interaction term between Gloss and Test format was added.

The analysis of immediate posttests revealed that the main effects of Gloss and Test format were significant, $F(21.7) = 67.5$, $p < .001$, $F(9.46) = 6.53$, $p = .017$, respectively, but the interaction was non-significant, $F(8.34) = 1.17$, $p = .356$. Subsequent multiple comparisons using the model without the interaction term showed that recognition tests led to a 24.8% higher ES than recall tests ($b = 0.248$, $p = .003$). Other tests (i.e., VKS and gap-filling tests) led to a 20.5% higher mean ES than recall tests, which was significant ($b = 0.205$, $p = .045$). There was no significant difference between other tests and recognition tests ($p = .589$).

The analysis of delayed posttests revealed that the interaction between Gloss and Test format was significant, $F(12) = 4.41$, $p = .037$, indicating that glossing effects differed depending on which tests were administered. Subsequent multiple comparisons using the model with interaction term showed that the glossing effect was 9.3% higher when measured with recognition tests compared to when measured with recall tests ($b = 0.093$, $p = .008$). There was no significant difference in gloss effects between other tests and recall tests ($p = .574$), or other tests and recognition tests ($p = .127$).

Table 2
The Overall Effect of Glossed-reading: Separately for Each Test Format

		Immediate					Delayed				
Test Format	Gloss	*k*	*n*	Mean ES (%)	CI	*p*	*k*	*n*	Mean ES (%)	CI	*p*
Recall	Gloss	20	57	28.4	18.2, 38.7	< .001	17	35	21.2	14.6, 27.9	< .001
	Non-gloss	13	16	14.1	5.8, 22.5	.002	13	16	13.3	6.4, 20.2	< .001
Recognition	Gloss	28	84	54.6	45.4, 63.9	< .001	25	60	42.9	35.6, 50.3	< .001
	Non-gloss	20	25	33.8	25.0, 42.6	< .001	20	24	25.6	18.4, 32.9	< .001
Other	Gloss	6	13	51.6	34.7, 68.6	< .001	8	18	26.8	21.4, 32.2	< .001
	Non-gloss	4	4	26.8	-0.9, 54.5	.055	6	7	12.0	-1.1, 25.2	.066

Note. *k* = number of studies, *n* = number of effect sizes, CI = 95% confidence interval, Mean ES = weighted mean effect sizes converted into a percentage from a proportion for the sake of interpretability. *p* = *p*-value for significant test.

Research Question 2: What are the Relative Effects of Gloss Types?

The number of studies including non-glossed conditions was 30, marginal glosses = 26, multiple-choice glosses = 13, hyperlinked glosses = 12, in-text glosses = 4, glossaries = 3, and interlinear glosses = 1. Test format was also inserted as a covariate to control for the potential bias due to the test format (e.g., less effective gloss types might not have been measured with more demanding tests).

The analysis of immediate posttests showed that different gloss types uniquely contributed to vocabulary learning. Table 3 shows the difference between each gloss type and the non-glossed condition. Multiple-choice led to the greatest gain, followed by marginal, hyperlinked, glossaries, interlinear, and in-text glosses in that order. While multiple-choice, hyperlinked, marginal, and interlinear glosses led to significantly greater learning compared to non-gloss ($p < .05$) and intext-glosses approached statistical significance ($p = .055$), glossaries did not reach the statistical significance ($p = .134$). Subsequent multiple comparisons revealed that multiple-choice glosses contributed to significantly higher scores than in-text glosses ($b = 0.142$, $p = .026$), interlinear ($b = 0.092$, $p = .026$), and marginal ($b = 0.074$, $p = .035$). Interlinear glosses were significantly higher than in-text glosses ($b = 0.050$, $p = .044$) and significantly lower than marginal glosses ($b = 0.142$, $p = .023$), but note that only four ESs from Zarei and Hasani (2011) accounted for interlinear glosses.

The analysis of delayed posttests revealed a slightly different pattern for each gloss type. The most effective gloss type on delayed posttests was multiple-choice, followed by hyperlinked, marginal glosses, glossaries, and in-text glosses in that order. In-text glosses and glossaries did not significantly differ from non-gloss ($p = .147$, $p = .412$, respectively). Multiple comparisons did not find any significant differences between each of the other gloss types.

Table 3
The Learning Gain for Each Gloss Type Compared to the Non-Glossed Condition

	Immediate					Delayed				
	k	*n*	Mean ES difference (%)	CI	*p*	*k*	*n*	Mean ES difference (%)	CI	*p*
MC	12	31	25.2	18.5, 31.8	< .001	12	21	15.6	9.0, 22.3	< .001
Hyperlinked	11	35	18.4	5.9, 30.9	.009	11	33	15.2	3.1, 27.3	.020
Marginal	25	69	17.8	13.5, 22.0	.001	21	50	12.8	9.6, 16.0	< .001
Glossaries	2	3	17.4	-27.7, 62.5	.134	3	5	10.4	-9.9, 30.6	.147
Interlinear	1	4	16.0	8.5, 23.5	.004	0	0	-	-	-
In-text	4	12	11.0	-0.4, 22.4	.055	3	4	6.5	-19.0, 32.1	.412

Note. *k* = number of studies, *n* = number of effect sizes, CI = 95% confidence interval, Mean ES difference (%) = mean effect size differences between each gloss type and the non-glossed condition converted into a percentage. *p* = *p*-value for significant test.

Research Question 3: What are the Relative Effects of Gloss Languages?

Among the included studies, 31 included the L1 gloss condition, 25 included L2 glosses, 3 included L1 plus L2 glosses, and 30 studies included non-glossed conditions. Table 4 shows the difference between each gloss language and non-glossed conditions. All gloss languages led to significantly greater learning gains compared to the non-glossed condition. L1 plus L2 led to the highest gain followed by L1, and L2 glosses in that order. Multiple comparisons revealed that L1 glosses contributed to 4.0% higher gains than L2 glosses (b = 0.040, p = .075). There were no significant differences between L1 plus L2 glosses and L1 (p = .367) and between L1 plus L2 glosses and L2 (p = .139).

The analysis of delayed posttests showed a similar trend as the results of immediate posttests. Every gloss language led to significantly greater learning gains than the non-glossed condition. Although the order of effectiveness was slightly different from the results of immediate posttests (L1, L1 plus L2, and L2 glosses), multiple comparisons found the same results as immediate posttests; that there was a statistical significance between L1 and L2 glosses (b = 0.052, p = .048). There were no significant differences between L1 plus L2 glosses and L1 (p = .962) and between L1 plus L2 glosses and L2 (p = .138).

Table 4
The Learning Gain for Each Gloss Language Compared to the Non-Glossed Condition

	Immediate					Delayed				
	k	*n*	Mean ES difference (%)	CI	*p*	*k*	*n*	Mean ES difference (%)	CI	*p*
L1 plus L2	3	4	23.7	9.0, 38.5	.021	3	4	16.0	7.1, 24.8	.017
L1	30	77	20.3	15.4, 25.3	< .001	27	56	16.1	11.4, 20.7	< .001
L2	23	65	16.3	10.9, 21.8	< .001	21	47	10.9	6.0, 15.8	< .001

Note. *k* = number of studies, *n* = number of effect sizes, CI = 95% confidence interval, Mean ES difference (%) = mean effect size differences between each gloss language and the non-glossed condition converted into a percentage. *p* = *p*-value for significant test.

Research Question 4: What are the Relative Effects of Gloss Modes?

Forty-two studies included the textual gloss condition, 5 studies included pictorial glosses, 2 studies included audio glosses, and 30 studies included non-glossed conditions. Table 5 shows the difference between each gloss mode and no-gloss. For immediate posttests, every gloss mode led to significantly greater learning gains compared to the non-glossed condition. Auditory glosses led to the highest ES, followed by pictorial glosses, then textual glosses. However, this result should be interpreted with caution because the ESs of auditory glosses came from only two studies (i.e., Rassaei, 2017; Sadeghi & Ahmadi, 2012). Multiple comparisons did not find any significant differences across different gloss modes.

The analysis of delayed posttests showed a similar trend as the results of immediate posttests (Figure 6). Auditory glosses led to the greatest ES; however, it did not reach statistical significance when compared with the non-glossed condition ($p = .108$). This may be due to the small sample size (only 2 studies accounted for auditory glosses). Pictorial and textual glosses led to greater learning gains than no-gloss ($p = .023$, $p < .001$, respectively). Multiple comparisons found that auditory glosses were marginal significantly higher than pictorial glosses ($b = 0.221$, $p = .091$). No significant differences were found between textual and auditory glosses ($p = .177$), or between textual and pictorial glosses ($p = .396$).

Table 5
The Learning Gain for Each Gloss Mode Compared to the Non-Glossed Condition

	Immediate					Delayed				
	k	*n*	Mean ES difference (%)	CI	*p*	*k*	*n*	Mean ES difference (%)	CI	*p*
Auditory	2	3	36.3	1.9, 70.7	.047	2	3	36.8	-40.9, 114.5	.108
Pictorial	4	8	25.4	4.7, 46.1	.029	3	6	15.0	5.0, 25.1	.023
Textual	39	143	18.0	13.4, 22.5	< .001	36	104	12.7	9.1, 16.4	< .001

Note. *k* = number of studies, *n* = number of effect sizes, CI = 95% confidence interval, Mean ES difference (%) = mean effect size differences between each gloss mode and the non-glossed condition converted into a percentage. *p* = *p*-value for significant test.

Research Question 5: To What Extent do Text and Learner Characteristics Moderate Glossing Effects?

Text characteristics. Regarding the L2 learner targeted material variable, in 29 studies, participants read texts that targeted L2 learners. In 5 studies, participants read texts that were written for native speakers. Neither of the analyses on immediate nor delayed posttests found any significant main effect (p = .229, p = .137, respectively) or interaction with glossing (p = .223, p = .540, respectively). This indicates that the audience that material was targeted for was not clearly associated with learning gains or glossing effects.

As for comprehensibility of texts, 16 studies reported participants' reading comprehension scores. After excluding six ESs from Jacobs et at. (1994) identified as outliers, 140 ESs were included into the analysis.[2] The analysis of immediate posttests showed that Comprehensibility was significantly associated with ESs (b = 0.004, p = .018) while the effect of glossing was controlled. This indicates that the estimated learning gain increases by 4% as the comprehension test score increases by 10% even when the effect of glossing is controlled. The interaction between Comprehensibility and Glossing was not significant (p = .437), which implies that reading comprehension and Glossing independently influenced vocabulary learning from reading (i.e., glossing may promote vocabulary learning regardless of how well learners comprehend texts). The same trend was found on delayed posttests; Comprehensibility was significantly associated with ESs (b = 0.004, p = 0.09) while the effect of glossing was controlled, and their interaction was not significant (p = .252).

For Target word percentage, the mean target word percentage was 2.7% (SD = 1.8%, Min = 0.6%, Max = 7.7%). Neither of the analyses on immediate nor delayed posttests found any significant main effect (p = .938, p = .898, respectively) or interaction with glossing (p = .716, p = .866, respectively).

For Text type, 18 studies used expository texts, and 17 studies used narrative texts. Neither of the analyses on immediate nor delayed posttests found any significant main effect (p = .472, p = .800, respectively) or interaction with glossing (p = .139, p = .171).

For CALL use, in 28 studies, participants read texts on a computer screen, while in 16 studies, participants read texts on paper. Neither analysis on immediate nor delayed posttests found any significant main effect (p = .614, p = .491, respectively) or interaction with glossing (p = .942, p = .850, respectively).

Learner characteristics. Out of 42 studies, 25 studies reported learners' L2 proficiency levels: 9 studies recruited beginner learners, 16 studies involved intermediate learners, and 3 studies recruited advanced learners. Studies reported participants' L2 proficiency based on different references. Eight studies (32%) referred to participants' English proficiency tests (e.g., International English Language Testing System [or IELTS], Oxford Placement Test, Test of English for International Communication [or TOEIC]), two studies (8%) used cloze tests, two studies (8%) looked at the context that participants were in (i.e., textbook used in the classroom, school administration's report), two studies (8%) looked at English proficiency tests and vocabulary size tests, and one study (4%) only used vocabulary size tests. The rest of the studies (10 studies, 40%) did not report how they determined proficiency.

The analyses of immediate and delayed posttests found that there was no significant main effect (p = .262, p = .331, respectively), suggesting that overall vocabulary learning gains did not significantly differ based on the how proficient students were. An interaction was significant on

immediate posttests (p = .048), showing that the glossing effect on vocabulary learning was moderated by learners' proficiency. Subsequent multiple comparisons of immediate posttests revealed that intermediate learners benefitted from glosses significantly more than beginners (b = 0.125, p = .026) and advanced learners (b = 0.187, p = .055). There was no significant difference between beginner and advanced learners (p = .395). The interaction was not significant on delayed posttests (p = .145); however, the same trend was observed, intermediate learners benefitted from glosses the most, followed by the beginners and advanced learners.

Do L2 proficiency levels benefit differently from different gloss languages? Because proficiency was not always reported, L1 plus L2 glosses suffered from a small number of studies (i.e., less than three) for each proficiency level. Hence, we focused on the L1 glosses and L2 glosses. For L1 glosses, 8 studies included beginners, 11 studies included intermediate level learners, and 3 studies included advanced learners. For L2 glosses, 8 studies included beginners, 11 studies included intermediate learners, and 3 included advanced learners. For non-gloss, 6 studies included beginners, 11 included intermediate learners, and 3 included advanced learners.

The analyses of immediate and delayed posttests did not find any significant interaction between gloss languages and proficiency (p = .430, p = .340, respectively), suggesting that L1 glosses yielded greater learning than L2 glosses, regardless of participants' L2 proficiency. However, this result should be interpreted with caution because a small number of studies included advanced learners and only a few studies compared different languages with students at different proficiency levels.

For participants' institutional levels, 31 studies recruited university students, 6 studies included language school students, and 5 studies recruited secondary school students. The analyses of immediate and delayed posttests found that there were no significant main effects (p = .398, p = .914, respectively) or interactions (p = .242, p = .150, respectively). However, there was a trend on the delayed posttests that language school students benefitted more from glossing more than secondary school (b = 0.183, 95% CI [0.338, 0.273], p = .030) and university students (b = 0.170, 95% CI [0.396, -0.055], p = .088).

Table 6
Moderator Analysis for Text and Learner Characteristics on Immediate Posttests

			Main effect		Interaction effect	
Variables	*k*	*n*	Coef. [CI]	*p*	Coef. [CI]	*p*
1. Text variables						
(1) L2 learner targeted material				.229		.223
A. Native speakers	5	23	-ref.-		-ref.-	
B. L2 learners	27	124	.110 [-.093, .314]		.080 [-.073, .233]	
(2) Comprehensibility				.016		.341
A. Comprehension test percentage	15	83	.004 [.001, .007]		.002 [-.004, .008]	
(3) Target word percentage				.938		.716
A. Target word percentage	30	151	-.002 [-.044, .040]		.006 [-.035, .048]	
(4) Text Type				.472		.171
A. Expository	18	81	-ref.-		-ref.-	
B. Narrative	14	65	.043 [-.078, .165]		.076 [-.036, .189]	
(5) CALL use				.614		.942
A. No-CALL	26	135	-ref.-		-ref.-	
B. CALL	15	64	.025 [-.077, .127]		-.004 [-.117, .109]	
2. Learner Characteristics						
(1) Proficiency				.262		.048
A. Beginner	9	46	-ref.-		-ref.-	
B. Intermediate	13	67	.146 [-.045, .338]		.125 [-.003, .253]	
C. Advanced	3	19	.024 [-.366, .414]		-.062 [-.233, .110]	

(2) Institutional Level				.398		.242
A. Secondary School	4	15	-ref.-		-ref.-	
B. University	30	163	.031 [-.231, .294]		-.039 [-.210, .131]	
C. Language School	5	21	.136 [-.124, .397]		.108 [-.058, .273]	

Note. *k* = number of studies. *n* = number of ESs. -ref.- = reference level. Coef. = estimated coefficient. *p* = *p*-value for significant test. CI = 95% confidence interval. Main effect refers to whether the moderator variable was related with the relative learning gains regardless of gloss provision, and Interaction effect refers to whether the moderator variable mediated the glossing effect (i.e., to what extent the mean learning gain through glossed reading differed from non-glossed reading).

Table 7
Moderator Analysis for Text and Learner Characteristics on Delayed Posttests

			Main effect		Interaction effect	
Variables	*k*	*n*	Coef. [CI]	*p*	Coef. [CI]	*p*
1. Text variables						
(1) L2 learner targeted material				.137		.540
A. Native speakers	4	16	-ref.-		-ref.-	
B. L2 learners	26	107	.142 [-.070, .354]		.021 [-.086, .128]	
(2) Comprehensibility				.009		.252
A. Comprehension test percentage	14	57	.004 [.001, .007]		.004 [-.004, .011]	
(3) Target word percentage				.898		.866
A. Target word percentage	28	126	-.003 [-.049, .044]		.003 [-.038, .044]	
(4) Text Type				.800		.139
A. Expository	15	59	-ref.-		-ref.-	
B. Narrative	16	74	-.014 [-.129, .101]		.068 [-.024, .159]	
(5) CALL use				.491		.850
A. No-CALL	23	97	-ref.-		-ref.-	
B. CALL	15	63	.040 [-.081, .162]		.009 [-.087 .105]	
2. Learner Characteristics						
(1) Proficiency				.331		.145
A. Beginner	7	26	-ref.-		-ref.-	
B. Intermediate	14	58	.126 [-.165, .417]		.132 [.035, .229]	
C. Advanced	2	11	-.072 [-.690, .546]		-.041 [-.216, .135]	

(2) Institutional Level				.914		.150
A. Secondary School	5	18	-ref.-		-ref.-	
B. University	27	119	.029 [-.152, .210]		.012 [-.055, .079]	
C. Language School	4	23	.043 [-.187, .273]		.183 [.027, .338]	

Note. *k* = number of studies. *n* = number of ESs. -ref.- = reference level. Coef. = estimated coefficient. *p* = *p*-value for significant test. CI = 95% confidence interval. Main effect refers to whether the moderator variable was related with the relative learning gains regardless of gloss provision, and Interaction effect refers to whether the moderator variable mediated the glossing effect (i.e., to what extent the mean learning gain through glossed reading differed from non-glossed reading).

Discussion and Conclusion

The current meta-analysis revealed that, on average, learners reading texts with glosses learned 45.3% of the unknown words on immediate posttests, and 33.4% on delayed posttests. These rates were significantly higher than learners who read texts without glosses (26.6% for immediate and 19.8% for delayed posttests). Glossing was found to contribute to vocabulary learning across all types of tests (i.e., recognition, recall, other). Glossing effects on delayed posttest scores were 9.3% smaller when measured with recall tests compared to when measured with recognition tests. This result suggests that learners may quickly become unable to recall target words they learned with glosses but most of them remain recognizable to a greater extent. These findings support earlier studies demonstrating that word recall is more difficult to acquire than recognition (González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004). In addition to word recognition and recall, knowledge required for other tests (VKS and gap-filling tests) was improved by glossing, indicating that glossed reading not only increases L2 learners' form-meaning connection but might also enhance knowledge required to use words.

Different Approaches to Glossing

Gloss type. The analysis of gloss type revealed that multiple-choice glosses were the most effective gloss type (mean ESs were 25.2% and 15.6% higher than non-glossed reading on immediate posttests and delayed posttests, respectively). Hyperlinked (18.4%, 15.2%), Marginal (17.8%, 12.8%), and interlinear glosses (16.0% on immediate posttest) led to somewhat similar effects. The effectiveness of glossaries (17.4%, 10.4%) and in-text glosses (11.0%, 6.5%) were not so clear; when compared to non-glossed reading, no significant difference was found on either immediate or delayed posttests.

The advantage of multiple-choice glosses over the other gloss types was distinct especially on immediate posttests. The characteristics of multiple-choice glosses, where learners have to read sentences to select the appropriate choice, which may have led to deeper processing and greater learning gains. The superiority of multiple-choice glosses may be explained by deeper processing (Hulstijn, 1992; Rott, 2005); the multiple-choice glosses require learners to read the sentence carefully and select the appropriate gloss among options. This extra processing may have strengthened the form-meaning connections of words. This finding is also in line with theories of vocabulary learning, mental effort hypothesis (Hulstijn, 1992) and involvement load hypothesis (Laufer & Hulstijn, 2001). Additionally, learners' skipping behavior may also explain the effect of multiple-choice glosses. Many studies found that students often ignore unknown words even when glosses were provided (e.g., Boers, Warren, Grimshaw, & Siyanova-Chanturia, 2017; Hulstijn et al., 1996; Warren et al., 2018). In contrast, in multiple-choice gloss conditions, students are told to look at each gloss to make a choice, so glossed words are not skipped. This enforced processing may lead to a greater learning rate.

Several factors potentially influence the effectiveness of multiple-choice glosses. For example, numbers of options may impact the degree to which learners evaluate each option for the sentence before selecting the best option (Hulstijn, 1992). Too many options might lead to more wrong selections and less learning, while too few options might hinder learning as learners easily choose the best option without deepening the processing of words. Additionally, the difficulty involving in selecting appropriate glosses may also be influenced by the similarities of options and participants' L2 proficiency levels. Lastly, some studies examined whether providing feedback for learners' selection enhances learning; however, the results are so far

inconsistent (Nagata, 1999; Yoshii, 2013). Multiple-choice glosses potentially hinder comprehension of the text (Martínez-Fernández, 2008; however, see also Rott, 2005). Studies directly looking at factors influencing the effectiveness of multiple-choice glosses are scarce, warranting further research exploring how multiple-choice glosses should be implemented.

The second most effective group of gloss types—i.e., marginal and hyperlinked—may benefit from their location. They are neither too close to nor too far from the target word. Marginal and hyperlinked glosses are close enough for learners to quickly check meanings of target words, which may provide more opportunities to make form-meaning links between unknown words and their meanings. In contrast, glossaries tend to be located far from target words; learners usually have to look at the end of the reading material to check a glossary. This might discourage learners from checking, potentially decreasing the frequency of consulting a glossary. In-text glosses could be too close to target words, as in-text glosses are provided right after each target word. Because learners can easily comprehend the message of the sentence just by looking at an in-text gloss, the target word does not necessarily have to be processed (Watanabe, 1997). It is reasonable to speculate that learners might have ignored glosses or target words given that all included studies in this meta-analysis set up incidental vocabulary learning conditions (i.e., participants were not told to learn target words) where comprehension of the text was the main purpose of reading. Participants may, therefore, have had no reason to check all glosses unless they had difficulty comprehending a text.

Only one study included interlinear glosses, which makes it difficult to draw a firm conclusion. Its effectiveness reached the statistical significance; however, the mean effect size was at the fifth place out of six. Similarly, the number of studies including glossaries and in-text glosses was relatively small (2-4). This could be because researchers would be more likely to hesitate to investigate the less effective gloss types based on their personal experience or intuitive reasoning. To arrive at a more robust conclusion, it would be useful to conduct more studies to examine the relative effectiveness of gloss types that are reported as less effective in this meta-analysis as well as ones reported as more effective.

Gloss language. The analysis of gloss language showed that L1 glosses and L1 plus L2 glosses led to similar learning gains. Reading with L2 glosses led to the smallest learning gain. This trend was consistent across immediate and delayed posttests. These findings indicate that unknown target words are more easily learned in glosses with L1 translations compared to L2 definitions or synonyms in general. Newly learned L2 words are typically mapped onto their L1 translation (Clenton, 2015; Jiang, 2002; Kroll & Stewart, 1994; Kroll, Van Hell, Tokowicz, & Green, 2010). Since L1 words tend to be more familiar to learners than L2 words, connections between L1 words and unknown L2 words can be more easily established compared to connections between L2 words and unknown L2 words (Choi, 2016). Students' skipping behavior could also explain the superiority of providing L1 glosses. Bell and LeBlanc (2000) found that participants more frequently consulted L1 glosses than L2 glosses. Boers et al. (2017) and Warren et al. (2018) tracked learners' eye-movements while reading glossed tests and found that unknown words annotated with L2 textual glosses were about twice as frequently ignored than pictorial glosses or glosses combining L2 textual and pictorial glosses. These findings point to the possibility that looking up unknown words in L2 glosses can be demanding and demotivate students from using glosses while reading. L1 glosses, in contrast, may be looked up more frequently than L2 glosses and lead to greater learning gains.

Regarding L2 proficiency, we did not find any support for the hypothesis that students with higher proficiency levels benefit more from L2 glosses than those with lower proficiency levels. There are two possible explanations for this. The first possibility is that the proficiency of participants may not have been high enough to benefit from L2 glosses as much as L1 glosses. Another possibility relates to types of studies accounting for advanced learners in this meta-analysis. Among the total of 3 studies, except for Choi (2016), none of the studies compared L1 and L2 glosses. In order to confirm the potential interaction between proficiency and gloss languages, more research recruiting advanced L2 learners is required.

It is probably worth emphasizing that although the results suggested an advantage of L1 glosses over L2 glosses, the effect of L2 glosses was clearly observed when compared to reading without glosses. Given the fact that L2 glosses are very useful especially for contexts where each learner's L1 background differs, the effectiveness of L2 glosses should not be disregarded.

Gloss mode. The analysis of gloss mode did not find clear differences between the different modes of glossing. Interestingly, the mean ES for auditory glosses was larger than for textual glosses and pictorial glosses. However, the difference was not statistically significant. These comparisons have to be interpreted with caution because only two studies included auditory glosses (Rassaei, 2017; Sadeghi & Ahmadi, 2012). Both studies reported that auditory glosses led to greater learning than textual glosses. Rassaei (2017) argued that the advantage of auditory glosses could be due to the fact that glosses were provided in a different channel than the one was used for reading (i.e., textual), which potentially allowed learners to direct more attention to the glosses or pay attention to glosses while looking at glossed words. Further research to evaluate the effect of auditory glosses is warranted.

Text Characteristics

Moderator analyses of text characteristics found that comprehension was a significant variable; reading comprehension and glossing independently promoted vocabulary learning. This result indicates that when learners read texts that are easier to understand, they learn more vocabulary from them. The results also imply that glossing promotes vocabulary learning even with reading materials that are relatively difficult for learners to comprehend. However, given that comprehensibility enhances vocabulary learning, it may be important to provide texts at the appropriate level to learners to maximize vocabulary learning even when providing glosses.

Other text related variables, L2 learner target materials, text type, and CALL use, were not significantly related to ESs or glossing effects. Lack of clear effects of L2 learner target materials may indicate that learning gains or glossing effects do not differ so much by whether or not reading materials are written for native speakers or L2 learners as long as the difficulty level of them are appropriate for learners. Studies using materials written for native speakers may have included relatively advanced learners.

Lack of a clear advantage of CALL use seems counterintuitive at first glance. Taylor's series of meta-analyses (2006, 2009, 2013, & 2014) focused on the effects of glosses on L2 reading comprehension and revealed the advantage of CALL glosses over paper-based glosses. One reason is that CALL glosses might contribute to vocabulary learning in a different mechanism from reading comprehension. Another potential explanation is that this meta-analysis operationalized CALL use as whether reading material was presented on a computer screen (as opposed to paper) without considering how the glosses were provided (e.g., gloss type and language). Providing glosses in a CALL context does not necessarily lead to better vocabulary

learning compared to reading on paper, and the type and language of glosses may impact more on the effectiveness of glossing.

Learner Characteristics

The results showed that glossing effects differed in relation to L2 proficiency levels. The intermediate students benefitted more from glossing than beginner and advanced learners. The superiority of gloss effects for intermediate students over beginners could be explained by the possibility that students with higher proficiency levels were able to utilize glosses more effectively. The smaller glossing effect for advanced learners than intermediate learners might be due to inferencing ability. Advanced learners may have had better ability to guess the meanings of unknown words while reading, reducing their need to focus on the glosses. Given that consulting glosses directs learners' attention away from reading, advanced learners might have more frequently ignored glosses while inferring word meanings in context.

The participants' institutional levels were not significant; however, there was a trend that students at language schools benefitted more from glossing than students at secondary schools. This could be explained by students' motivational factor. Students at language schools might be more motivated to learn a target language compared to students at secondary schools or universities, and they might have paid more attention to unknown words glossed in the text while thinking that reading is for the sake of learning the language.

Limitations and Future Directions

Because the current meta-analysis only focused on learning gains, it is not clear why specific gloss formats led to greater learning than other formats. To expand on earlier studies, it would be helpful to look into the process of learning while reading glossed materials (e.g., Rott, 2005). For example, recent studies utilizing eye-tracking technology address learners' behavior while reading with textual glosses, pictorial glosses, and multimodal glosses (Boers, Warren, Grimshaw, et al., 2017; Warren et al., 2018). Future research looking at learners' cognitive processes while learning using eye-tracking and/or think-aloud protocols may further reveal how learners benefit from different glossing approaches and which condition maximizes the effectiveness of glossing.

The present study revealed several areas requiring more attention. First, it would be helpful to further examine the effectiveness of interlinear glosses and audio glosses as we found few studies investigating these gloss formats. Second, regarding the interaction between gloss language and L2 proficiency, future studies should recruit advanced learners to compare the effects of L1 and L2 glosses. Studies recruiting learners of different proficiencies to investigate the relative effectiveness of different gloss languages may also allow a more direct and accurate examination. Third, clear reporting of L2 proficiency based on specific standardized references (e.g., Common European Framework of Reference for Languages or American Council on the Teaching of Foreign Languages Proficiency Guidelines) may allow future meta-analyses to more accurately investigate the relationship between treatment and proficiency. Fourth, most of the included studies focused on form-meaning mapping. Although this is common in vocabulary research (Uchihara et al., 2019), exploration of depth of vocabulary knowledge (e.g., grammatical functions, collocations, and associations) may further reveal how vocabulary knowledge develops from glossed reading (Webb, 2007; see also Yanagisawa & Webb, 2019 for a review of various approaches to measuring depth of vocabulary knowledge). Lastly, the effects of glosses may be influenced not only by each factor (e.g., gloss formats, characteristics of

learners and texts, and research methodology) but also by combinations of these factors. For example, it is difficult to treat L2 proficiency and text difficulty separately because researchers may select easier reading materials for less proficient learners and vice versa. This is something that meta-analysis cannot address. Hence, individual studies are needed to examine further complicated relationships among variables and how those influence students' vocabulary learning.

The process of this systematic analysis also found several methodological features requiring further attention. First, we found that test reliability statistics (e.g., Cronbach's α) on vocabulary test scores were not always reported; among the included 42 studies, only 16 studies (38.1%) reported test reliability statistics. Inconsistent reporting practice of test reliability was also observed by Uchihara et al. (2019) who found that only 24% of included studies reported this. Test reliability measures not only indicate the reliability of the test but also help authors identify idiosyncratic target items.

Second, while most of the studies (39 studies, 92.9%) recruited Foreign Language (FL) students as participants, only a few studies (3 studies, 7.1%) recruited Second Language (SL) students. Different learning contexts might influence students' attitude, strategies while reading, and familiarity with certain approaches to glossing, all of which potentially results in different effects of glossed reading. Future research should investigate glossing effects in different contexts, especially in SL contexts.

Lastly, we found that occasionally reading materials, test formats, and directions to participants were not so clearly described in studies. This prohibits exploring other potential moderator variables such as whether glossed words were underlined/highlighted, which language was used for testing, and whether participants were explicitly told to look up unknown words while reading. Providing materials (texts, tests) in the study as an appendix may enhance the clarity of studies. Following a recent trend in the field recommending open data and open materials (e.g., Marsden, Trofimovich, & Ellis, 2019), we would like to encourage researchers to release their materials (e.g., texts, test formats) and results dataset making them publicly available if possible. This may make future replications easier and more vigorous, as well as enhancing the transparency of the research. Having access to open materials and datasets helps future meta-analyses to provide a clearer picture of the effects of glossed reading, as well as make more accurate and robust estimations by taking advantage of individual participant data (see e.g., Cooper & Patall, 2009).

References

Note. The full reference list of the studies included in the meta-analysis is available in On-line Supplementary Materials Appendix 1.

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

AbuSeileek, A. F. M. (2013). Hypermedia annotation presentation: Learners' preferences and effect on EFL reading comprehension and vocabulary acquisition. *CALICO Journal*, *25*(2), 260–275.

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on l2 vocabulary acquisition: A comparative study. *Language Learning & Technology*, *5*, 202–232.

Azari, F. (2012). Review of effects of textual glosses on incidental vocabulary learning. *International Journal of Innovative Ideas*, *12*(2), 13–24.

Azari, F., Abdullah, F. S., Heng, C. S., & Hoon, T. B. (2012, July). Effects of glosses on vocabulary gain and retention among tertiary level EFL learners. Retrieved from https://eric.ed.gov/?id=ED533228

Bell, F. L., & LeBlanc, L. B. (2000). The language of glosses in L2 reading on computer: Learners' preferences. *Hispania*, *83*(2), 274–285. https://doi.org/10.2307/346199

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

Bowles, M. A. (2004). L2 glossing: To CALL or not to CALL. *Hispania*, *87*(3), 541. https://doi.org/10.2307/20063060

Cheng, Y. H., & Good, R. L. (2009). L1 glosses: Effects on EFL learners' reading comprehension and vocabulary retention. *Reading in a Foreign Language*, *21*(2), 119–142.

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

Choi, S. (2016). Effects of L1 and L2 glosses on incidental vocabulary acquisition and lexical representations. *Learning and Individual Differences*, *45*, 137–143. https://doi.org/10.1016/j.lindif.2015.11.018

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/10.2307/328635

Clenton, J. (2015). Testing the Revised Hierarchical Model: Evidence from word associations. *Bilingualism: Language and Cognition*, *18*(1), 118–125. https://doi.org/10.1017/S136672891400008X

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*(2), 165–176. https://doi.org/10.1037/a0015565

Day, R. R., Omura, C., & Hiramatsu, M. (1992). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, *7*, 541–551.

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*. https://doi.org/10.1111/lang.12296

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fang, S. (2009). *Chinese gloss or English gloss: Which is more effective for incidental vocabulary acquisition through reading?* Kristianstad University College. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:hkr:diva-1045

González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, Advance online publication. https://doi.org/10.1093/applin/amy057

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *International Review of Applied Linguistics in Language Teaching; Heidelberg*, *41*(2), 87–106.

Holley, F. M., & King, J. K. (1971). Vocabulary glosses in foreign language reading materials. *Language Learning*, *21*(2), 213–219. https://doi.org/10.1111/j.1467-1770.1971.tb00060.x

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207–223.

Hu, H. M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 113–125). London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-12396-4_11

Hulstijn, 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: Cambridge University Press.

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*(3), 327–339. https://doi.org/10.2307/329439

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed). Thousand Oaks, CA: Sage.

Jacobs, G. M., Dufon, P., & Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, *17*(1), 19–28. https://doi.org/10.1111/j.1467-9817.1994.tb00049.x

Jalali, S., & Neiriz, A. (2012). Computer-based versus traditional L1 and L2 glosses, *2*(9), 188–217.

Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *24*(04), 617–637.

Ko, M. H. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly*, *46*(1), 56–79. https://doi.org/10.1002/tesq.3

Ko, M. H. (2017). The relationship between gloss type and L2 proficiency in incidental vocabulary learning. *The Modern English Society*, *18*(3), 47–69. https://doi.org/10.18095/meeso.2017.18.3.03

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Kost, C. R., Foss, P., & Lenzini, 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., Hell, J. G. V., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, *13*(3), 373–381. https://doi.org/10.1017/S136672891000009X

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

Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability'. *Journal of Research in Reading*, *15*(2), 95–103. https://doi.org/10.1111/j.1467-9817.1992.tb00025.x

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, *21*(2), 32–51.

Lee, H., Warschauer, M., & Lee, J. H. (2018). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*. https://doi.org/10.1093/applin/amy012

Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, *16*(1), 33–42. https://doi.org/10.1177/003368828501600103

Marsden, E., Trofimovich, P., & Ellis, N. (2019). Extending the reach of research: Introducing open accessible summaries at language learning. *Language Learning*, *69*(1), 11–17. https://doi.org/10.1111/lang.12337

Martínez-Fernández, A. (2008). Revisiting the involvement load hypothesis: Awareness, type of task and type of item. In M. A. Bowles, R. Foote, S. Perpiñán, & R. Bhatt (Eds.), *Selected Proceedings of the 2007 Second Language Research Forum* (pp. 210–228). Somerville, MA: Cascadilla Proceedings Project.

Miyasako, N. (2002). Does text-glossing have any effects on incidental vocabulary learning through reading for Japanese senior high school students? *Language Education & Technology*, *39*, 1–20. https://doi.org/10.24539/let.39.0_1

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mohsen, M. A., & Balakumar, M. (2011). A review of multimedia glosses and their effects on L2 vocabulary acquisition in CALL literature. *ReCALL*, *23*(02), 135–159. https://doi.org/10.1017/S095834401100005X

Nagata, N. (1999). The effectiveness of computer-assisted interactive glosses. *Foreign Language Annals*, *32*(4), 469–479. https://doi.org/10.1111/j.1944-9720.1999.tb00876.x

Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly*, *37*(4), 645–670. https://doi.org/10.2307/3588216

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd Edition). New York, NY: Cambridge University Press.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28.

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In *Advancing quantitative methods in second language research* (pp. 106–128). New York, NY: Routledge. Retrieved from https://www.routledge.com/products/9780415718349

Pustejovsky, J. (2018). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections (Version 0.3.1). Retrieved from https://CRAN.R-project.org/package=clubSandwich

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rassaei, E. (2017). Computer-mediated textual and audio glosses, perceptual style and L2 vocabulary learning. *Language Teaching Research*, Advance online publication. https://doi.org/10.1177/1362168817690183

Raudenbush, S. W. (2009). Analyzing effect sizes: Random effects models. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). New York, NY: Russell Sage Foundation.

Rott, S. (2005). Processing glosses: A qualitative exploration of how form-meaning connections are established and strengthened. *Reading in a Foreign Language*, *17*(2), 95–124.

Sadeghi, K., & Ahmadi, N. (2012). The effect of gloss type and mode on Iranian EFL learners' vocabulary acquisition. *Issues in Language Teaching*, *1*(1), 159–188.

Salehi, V., & Naserieh, F. (2013). The effects of verbal glosses on vocabulary learning and reading comprehension. *Asian EFL Journal*, *15*, 24–64.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

So, W. (2010). *Short-term and long-term retention of new words: Investigating the role of L1 glossing in vocabulary learning among Hong Kong ESL learners* (Unpublished master's thesis). The University of Hong Kong, Pokfulam, Hong Kong. Retrieved from http://hub.hku.hk/handle/10722/132270

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261–285. https://doi.org/10.2307/1170540

Taylor, A. (2006). The effects of CALL versus traditional L1 glosses on L2 reading comprehension. *CALICO Journal*, *23*(2), 309–318.

Taylor, A. (2013). CALL versus paper: In which context are L1 glosses more effective? *CALICO Journal; San Marcos*, *30*(1), 63–81. http://dx.doi.org/10.11139/cj.30.1.

Taylor, A. (2014). Glossing frequency and L2 reading comprehension: The influence of CALL glossing. *CALICO Journal*, *31*(3), 374–389. https://doi.org/10.11139/cj.31.3.374-389

Taylor, A. M. (2009). Call-based versus paper-based glosses: Is there a difference in reading comprehension? *CALICO Journal*, *27*(1), 147–160. https://doi.org/10.11139/cj.27.1.147-160

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599. https://doi.org/10.1111/lang.12343

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*(1), 3–25.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw*, *36*(3), 1–48.

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

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*, 1–24. https://doi.org/10.1017/S0272263118000177

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *19*(3), 287–307. https://doi.org/10.1017/S027226319700301X

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S., & Chang, A. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*(6), 667–686. https://doi.org/10.1177/1362168814559800

Webb, S., & Nation, I. S. P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, *16*, 1–10.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Place of publication not identified: Oxford University Press.

Xu, X. (2010). The effects of glosses on incidental vocabulary acquisition in reading. *Journal of Language Teaching and Research*, *1*(2), 117–120. https://doi.org/10.4304/jltr.1.2.117-120

Yanagisawa, A., & Webb, S. (2019). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 371–386). https://doi.org/10.4324/9780429291586-24

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*(2), 48–67.

Yeh, Y., Wang, C. W., & Tsing, N. (2003). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journa*, *21*(1), 131–144.
Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, *10*(3), 85–101.
Yoshii, M. (2013). Effects of gloss types on vocabulary learning through reading: Comparison of single translation and multiple-choice gloss types. *CALICO Journal*, *30*, 203–229.
Yoshii, M., & Flaitz, J. (2002). Second language incidental vocabulary retention: The effect of text and picture annotation types. *CALICO Journal*, *20*(1), 33–58.
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
Zarei, A. A., & Hasani, S. (2011). The effects of glossing conventions on L2 vocabulary recognition and production. *The Journal of Teaching Language Skills*, *3*(2), 209–233.
Zhao, T., & Ren, J. (2017). Incidental L2 lexical acquisition in reading: The role of L2-gloss frequency and learner proficiency. *The Language Learning Journal*, 1–17. https://doi.org/10.1080/09571736.2017.1349168

On-line Supplementary Materials
Appendix 1. References for the Included Studies
Appendix 2. Flowchart for inclusion and exclusion of the studies
Appendix 3. Basic Information about the Included Studies
Appendix 4. Detail of the analytic approach and Calculation Formula for ESs and *SDs*
Appendix 5. Additional Analyses

Notes

[1] We also coded methodological characteristics of studies, such as participant allocation methods (e.g., random allocation, intact-class allocation, and systematic allocation) and delayed posttest timing (i.e., number of days between the treatment and the delayed posttest) and analyzed their relationship with effect sizes. However, because these are not the current meta-analysis' focus, we did not include the results here. For those interested, please see the On-line Supplementary Materials Appendix 5.

[2] We conducted the analyses with and without outliers and confirmed the results show the same trend of the data.