How Effective Are Intentional Vocabulary-Learning Activities? A Meta-Analysis

STUART WEBB,[1] AKIFUMI YANAGISAWA,[2] AND TAKUMI UCHIHARA[3]
[1] *University of Western Ontario, Faculty of Education, 1137 Western Road, London, ON N6G 1G7, Canada Email: swebb27@uwo.ca*
[2] *University of Western Ontario, Faculty of Education, 1137 Western Road, London, ON N6G 1G7, Canada Email: ayanagis@uwo.ca*
[3] *University of Western Ontario, Faculty of Education, 1137 Western Road, London, ON N6G 1G7, Canada Email: takumi1356@gmail.com*

The present meta-analysis aimed to summarize the extent to which second language vocabulary is learned from the most frequently researched word-focused activities: flashcards, word lists, writing, and fill-in-the-blanks. One hundred effect sizes from 22 studies were included in meta-regression analyses and administered separately for the observations measured with meaning-recall and form-recall tests. The results revealed that the average percentage learning gains were 60.1% and 58.5% on meaning-recall and form-recall immediate posttests. These gains dropped to 39.4% and 25.1% on delayed meaning- and form-recall tests, respectively. These results suggest that learning through word-focused tasks is far from guaranteed. Moreover, the percentage learning gains among the different activities ranged from 18.4% to 77.0% on immediate posttests and from 23.9% to 73.4% on delayed posttests indicating that there is much variation in efficacy among the activities. Moderator analyses revealed that learners' place of study and direction of learning affected learning.

Keywords: vocabulary; meta-analysis; intentional learning; activity type; flashcards; writing; fill-in-the-blanks; second language acquisition

APPROACHES TO VOCABULARY LEARNING are often seen as belonging to an incidental–intentional dichotomy (e.g., Laufer, 2003; Nation, 2013; Schmitt, 2000, 2008; Webb, 2020; the terms intentional, deliberate, instructed, and explicit are often used synonymously in the literature). Meaning-focused approaches to learning such as reading, listening, and viewing (television and movies) are examples of incidental vocabulary-learning methods. Research shows that words can be learned incidentally through reading (e.g., Pellicer–Sánchez & Schmitt, 2010), listening to passages (e.g., van Zeeland & Schmitt, 2013) and songs (Pavia, Webb, & Faez, 2019), and watching television (Feng & Webb, 2020; Peters & Webb, 2018; Rodgers & Webb, 2020). Approaches that include an effort to attend to and learn words are viewed as being intentional vocabulary-learning methods. Research indicates that vocabulary can be learned through intentional activities such as flashcards (Nakata, 2008), word lists (Mondria & Wiersma, 2004), writing (Webb, 2005), and fill-in-the-blanks (Rott, 2012). The value in grouping activities according to these intentional and incidental labels is that the classification helps to differentiate between the relative effects of these two broad approaches to learning words. However, a limitation of classifying activities as incidental and intentional is that efficacy tends to be generalized across each category. For example, intentional learning activities are described as being most effective and providing the greatest chance that words will be learned (Schmitt, 2000). Schmitt (2008) stated that "intentional vocabulary learning almost always leads to greater

The Modern Language Journal, 0, 0, (2020)
DOI: 10.1111/modl.12671
0026-7902/20/1–24 $1.50/0

and faster gains, with a better chance of retention and of reaching productive levels of mastery than incidental vocabulary learning" (p. 341). There is justification for these statements, because the efficacy of intentional approaches to vocabulary learning is supported by studies showing that intentional approaches contribute to significantly greater vocabulary-learning gains than incidental approaches (e.g., Laufer, 2003, 2005). However, what remains unclear is the degree to which intentional and incidental study of vocabulary leads to learning, as well as the extent to which learning is consistent across different types of incidental and intentional vocabulary-learning activities.

There are many ways to intentionally learn second language (L2) words. Morgan and Rinvolucri (2004) described 118 activities designed to develop vocabulary knowledge, while Webb and Nation (2017) profiled 23 approaches to learning words that they viewed as being most effective. With so many ways of learning words, it is important to understand the extent to which different approaches are effective. Surprisingly, however, there is little clarity about the relative efficacy of different vocabulary-learning activities.

Meta-analysis provides greater transparency of research findings and the factors that affect learning. For example, there is often a great deal of variation in the amount of learning in a single learning condition. However, by accounting for factors that affect vocabulary learning such as activity type (e.g., Laufer & Shmueli, 1997), test format (Laufer & Goldstein, 2004), and learning direction (receptive or productive; e.g., Webb, 2009a, 2009b), meta-analytic research can provide a much clearer indication of the effects of each learning condition, as well as the factors that affect learning. The aim of the present study was to synthesize the findings of the most frequently researched word-focused activities: flashcards, word lists, writing, and fill-in-the-blanks. A second aim was to investigate the extent to which different factors such as participants' place of study (secondary school, university) and time on task moderate results. The research should provide greater transparency about the proportional gains that occur through completing common word-focused activities, as well as the degree to which these gains vary between activities. The results of this study should help to guide teachers and learners to optimize vocabulary learning by identifying the most effective and efficient activities.

DEFINING INTENTIONAL VOCABULARY LEARNING

Intentional vocabulary-learning activities can be defined in several ways. It might be tempting to define intentional vocabulary learning as conscious learning and incidental vocabulary learning as subconscious learning. However, this definition is problematic because conscious word learning likely varies between words and between learners within individual tasks. Intentional word learning can also be defined by whether participants know that they will be tested on their word learning (e.g., Hulstijn, 2001). This definition is frequently used in research in psychology and allows differentiation between incidental learning, where participants are unaware of a subsequent vocabulary test, and intentional learning, where they know they will be tested. Perhaps the most common and inclusive definition of intentional vocabulary learning is completing activities that are designed to promote word learning. Such activities clearly focus attention on the words to be learned. For example, crossword puzzles, word search, semantic mapping, word lists, word parts tables, and *Pictionary* all meet this definition. The present study adopts this final definition because it has ecological validity within the field of applied linguistics; inside and outside the language learning classroom, students complete tasks with the purpose of learning words. They are not always tested on what they learn, and if they are tested, they may often be without knowledge of a subsequent vocabulary test.

INTENTIONAL VOCABULARY-LEARNING GAINS

Despite the large number of ways to learn vocabulary through direct study, research has primarily focused on only four activities: flashcards, word lists, fill-in-the-blanks, and writing words in sentences and compositions. This is surprising, because other activities such as form–meaning matching, multiple-choice questions, and crosswords are found in most English language course books. Thus, we might expect to find research on the efficacy of a wider range of activities.

There are many studies of L2 vocabulary learning from flashcards perhaps due to the many factors (e.g., direction of learning, test format, number of target words) that researchers can control and investigate their effects in the paired associate paradigm (e.g., Nakata, 2016; Waring, 1997). Flashcards have generally been viewed as a highly

effective method of learning words (Nation, 2013). However, vocabulary-learning gains have been somewhat inconsistent. For example, studies have revealed percentage learning gains under 40% (e.g., Nakata, 2008, 2016; Rimrott, 2010) and over 80% (e.g., Nakata, 2016; Waring, 1997).

Studying vocabulary in word lists is another highly examined area (e.g., Mondria & Wiersma, 2004; Webb, 2009a, 2009b). Research suggests that learning from word lists is not as effective as learning from flashcards because retrieving the L2 forms or meanings of the target items has a positive effect on learning in the flashcards condition (Barcroft, 2007). There is a great deal of variation in the percentage learning gains in studies investigating learning from word lists. For example, Mondria and Wiersma (2004) found gains, for learners studying with word lists, of 53% on an immediate form-recall test and 12% on a delayed form-recall test 13 days later. They also found gains as high as 98% on an immediate meaning-recall test with gains on the corresponding delayed test at 48%. Nakata (2008) found gains of 82% on an immediate form-recall test, but gains of only 26% on that test 4 days later. Perhaps one reason for the variation in findings is that there are different ways in which learners may study word lists; some learners may cover the form or meaning of the words and try to retrieve the covered information in an approach similar to flashcards. Others, however, may simply look at both the forms and meanings of words together.

Writing words in sentences and compositions has also led to great differences in percentage learning gains. For example, Webb (2005) found that learners who wrote words in sentences had gains of 88% on an immediate meaning-recall test. Similarly, Javanbakht (2011) found that writing words in sentences led to gains of 84% on a meaning-recall test. In contrast, Pichette, de Serres, and Lafontaine (2012) found that participants who wrote words in sentences knew only 25% of those words on an immediate form-recall test, and only 11% of the words on the same test 1 week later. Similarly, Keating (2008) found that participants who wrote words in sentences knew only 21% of the words on a form-recall test.

The results of studies investigating word learning through completing fill-in-the-blanks exercises have also revealed variation in percentage learning gains. However, the gains have tended to be lower with maximum and minimum percentage learning gains of 66% (Ansarin & Bayazidi, 2016) and 7% (Rott, 2012), respectively. Percentage learning gains have tended to range between 10% and 40% (Hulstijn & Laufer, 2001; Keyvanfar & Badranghi, 2011; Tu, 2004).

VARIABLES THAT MAY AFFECT VOCABULARY LEARNING

Activity Type

Activities for learning vocabulary are often compared to determine which approach to learning is most effective (e.g., Laufer, 2003; Webb, 2005). For example, Laufer (2003) found that sentence completion, writing words in sentences, and writing words in compositions contributed to greater vocabulary-learning gains than encountering target words during reading.

Test Format

Research indicates that the use of different test formats affects the degree to which test takers are able to demonstrate vocabulary-learning gains (Laufer & Goldstein, 2004). Test takers tend to score highest on meaning-recognition formats (e.g., multiple-choice or matching formats that involve selecting the correct meanings for target words) followed by form-recognition (e.g., multiple-choice or matching formats that involve selecting the L2 forms that represent given meanings), meaning-recall (e.g., write the meaning of the given word), and form-recall (e.g., write the L2 word that corresponds with a given meaning) formats.

Level of L2 Proficiency

Several studies have indicated that more advanced learners may make greater incidental vocabulary-learning gains than less proficient learners (de Vos et al., 2018; Webb & Chang, 2015). This is intuitively logical because greater knowledge of the L2 should help students to understand and use language. The degree to which L2 proficiency may affect intentional word learning gains is rarely examined. However, a meta-analysis of studies with learners at different proficiency levels can shed more light on this variable.

Learning Direction

Several studies with bidirectional tasks have indicated that receptive and productive learning have different effects on the type of vocabulary knowledge gained (e.g., Waring, 1997; Webb, 2009a, 2009b). Transfer-appropriate processing

theory (Morris, Bransford, & Franks, 1977) suggests that the similarity between learning and testing conditions affects findings: Receptive learning may lead to greater gains in receptive knowledge, while productive learning may lead to greater gains in productive knowledge. Learning words in the productive direction using flashcards, in which students are presented with the meaning of words and must produce their L2 forms, leads to greater gains in productive vocabulary knowledge than receptive learning in which L2 words are encountered. In contrast, learning words in the receptive direction with flashcards (students are presented with the L2 forms of words and then produce their meanings) may lead to greater receptive vocabulary knowledge than productive learning. There are few explicit comparisons of other receptive and productive learning activities (for one exception, see Webb, 2005).

Time on Task

It is reasonable to assume that activities that take longer lead to greater learning than those that take less time. This is partially supported by research. Productive learning of flashcards takes longer and leads to greater gains in productive knowledge than receptive learning (e.g., Waring, 1997). In contrast, receptive learning of flashcards leads to greater gains in receptive vocabulary knowledge than productive flashcard learning, despite taking less time (e.g., Waring, 1997). Webb (2005) also found that a more time-intensive sentence-writing activity led to greater word learning than a shorter sentence-reading activity. However, when time was the same for these activities, the results changed, and the reading activity led to larger gains.

Design

The studies being investigated involve either between-participants designs in which different groups of participants learn the same target words in different conditions, or within-participants designs in which the same participants learn different target words in multiple conditions. The advantages of between-participants designs are that, since participants typically complete a single learning condition, they require less time and there is less likelihood of participant fatigue. Disadvantages are a need for a greater number of participants and the possibility that learner characteristics between groups affects gains. In contrast, within-participants designs ensure that learner characteristics will not affect findings. Additionally, within-participants designs tend to have greater statistical power by accounting for the variance related to each participant (e.g., Ansarin & Bayazidi, 2016; Larson–Hall, 2010). Therefore, the number of participants can be smaller to detect a significant effect compared to between-participants design. However, results may potentially be affected by participant fatigue. These design features have not been examined in earlier research on vocabulary learning.

Test Scoring

In test formats such as form-recall that require participants to write the L2 forms of target words, researchers may opt to score responses strictly. This involves only scoring words as correct when they are spelled correctly. Researchers may also score words for partial knowledge. This involves scoring words as correct when they are spelled correctly or spelled incorrectly but in such a way as to be able to demonstrate partial knowledge of the L2 forms of target words (e.g., Barcroft, 2007; Webb, 2005). For example, Nakata and Webb (2016) scored the following responses as correct for the target word apparition: "apparation," "appartion," and "applition."

Measurement of Prior Knowledge of Target Words

One methodological issue that researchers must struggle with is how to determine whether participants have prior knowledge of target words. Prior knowledge of target words is often measured with a pretest (e.g., Nakata, 2016). However, the use of a pretest–posttest design may often contribute to test effects in which participants gain knowledge of the target words through subsequent administrations of the test (e.g., Webb & Chang, 2015). Researchers may also adopt a design that involves measuring prior knowledge of target words with learners who have the same learner characteristics as the participants (e.g., Bao, 2015; Hulstijn & Laufer, 2001). This eliminates the possibility of a test effect but may lead to concerns that perhaps the knowledge of the matched learners is not the same as the actual participants. A third option is to use nonwords (e.g., *ancon*, *hodet*) as target words (e.g. Webb, 2009a, 2009b). This eliminates the need to measure prior knowledge because participants cannot know words that do not exist. However, it may lead some to question the ecological validity of the nonwords; would the gains for real words be the same as those for nonwords?

Number of Target Words

The number of target words in studies investigating intentional vocabulary learning varies between studies. For example, participants in one study attempted to learn six words (Ansarin & Bayazidi, 2016) while participants in another tried to learn 30 (Rimrott, 2010). There is some evidence that the number of target words does not affect learning gains (Nakata & Webb, 2016). However, increasing the number of target words in a task would likely increase the learning burden for participants.

THE PRESENT STUDY

The present meta-analysis focused on studies investigating four types of word-focused activities—namely, fill-in-the-blanks, writing activities, flashcards, and word lists—in L2 instructional settings. We exclusively analyzed quasi-experimental studies in which L2 students learned unfamiliar words in a single session and were subsequently tested. Previous meta-analyses (e.g., Huang, Willson, & Eslami, 2012; Won, 2008) often calculate effect sizes (ESs) based on the mean differences between experimental and control groups or between pretest and posttest scores (e.g., Cohen's *d*). However, in studies investigating word-focused activities, participants normally learn unknown L2 words through completing activities. These studies tend not to have a true control group or report pretest scores, making a calculation of ESs based on the mean differences difficult. To deal with this issue, the present meta-analysis computed absolute learning gains—that is, the proportion of the target words learned, and the number of target words learned per minute—as target ESs. These ESs may provide a clearer picture of the extent to which L2 students learn target words by engaging in word-focused activities. Accordingly, the following research questions were posed:

RQ1. To what extent is L2 vocabulary learned through completing word-focused activities?
RQ2. To what extent do L2 vocabulary-learning gains differ among word-focused activities?
RQ3. How many words are learned per minute through completing word-focused activities?
RQ4. To what extent does the number of words learned per minute vary through completing word-focused activities?
RQ5. What are the potential factors that moderate the effects of word-focused activities?

METHOD

Literature Search

Following In'nami & Koizumi (2010) and Plonsky & Oswald's (2015) guidelines, the following electronic databases were searched to identify studies that might be included in the meta-analysis: Education Resources Information Center (ERIC), PsycINFO, and Linguistics and Language Behavior Abstracts (LLBA). ProQuest Global Dissertations was also employed to identify unpublished PhD theses that might be included. Various combinations of keywords such as L2/foreign/second, vocabulary/word, learning/acquisition/retention, and activity/task/exercise, and the names of activities (e.g., writing, fill-in, flashcard/card, word list) were used to identify resources to include in the meta-analysis. Furthermore, given that studies investigating the Involvement Load Hypothesis (Laufer & Hulstijn, 2001) and Technique Feature Analysis (Nation & Webb, 2011) often examine word-focused activities, keyword search with involvement load hypothesis, involvement load, and technique feature analysis were also conducted. As a result, the titles and abstracts of 2,712 reports that appeared initially eligible for the meta-analysis were retrieved. The reference sections of primary publications concerning word-focused activities (Laufer, 2003, 2005) and L2 vocabulary learning (Nation, 2013), along with studies included in previous meta-analyses on word-focused activities (Huang et al., 2012; Won, 2008), were also used to identify relevant studies. All these publications were carefully examined in reference to the following selection criteria.

Criteria for Inclusion

The following 10 criteria for inclusion were used to evaluate the retrieved studies:

1. The study measured L2 vocabulary gains through completing activities in which unknown target words had to be attended to in order to complete the goals of the activities.
2. The study investigated the learning of specific sets of target vocabulary. Studies in which additional words beyond the target words were learned were excluded. This

is because the posttest scores and time on task from these studies may not represent accurate learning gains from one session of word-focused activities.
3. The study involved word-focused learning conditions that were likely to occur in the classroom (e.g., fill-in-the-blanks, writing sentences) or at home (e.g., flashcards, word lists). Laboratory-based studies that focused on the influence of specific factors on learning (e.g., influences of stimulus characteristics and background music on paired-associate learning; de Groot, 2006) or had limited amounts of time for learning (e.g., Barcroft, 2004) were excluded.
4. The study included one condition (or group) that was limited to only one type of activity. Studies that involved learning words through multiple activities were excluded (e.g., Groot, 2000: Laufer, 2006; Laufer & Shmueli, 1997).
5. The study treatment lasted up to 1 day. To attain a clear picture of learning and retention from engaging in one word-focused activity, studies with treatments lasting longer than 1 day were excluded (e.g., Azabdaftari & Mozaheb, 2012; Choi & Ma, 2015).
6. The study examined single word learning. Studies that investigated the learning of formulaic language (e.g., collocations) were excluded (e.g., Cao, 2013).
7. The study focused on individual learning. Studies involving participants learning in pairs or groups were excluded (e.g., Moonen et al., 2014).
8. The study used meaning-recall or form-recall tests. There were insufficient numbers of other test formats (i.e., form recognition, meaning recognition) to provide reliable data using these measurements. Accordingly, two studies that administered meaning recognition tests (i.e., Browne, 2003; Coomber, Ramstad, & Sheets, 1986) were excluded.
9. The study reported enough statistical information for the meta-analysis to be completed (i.e., mean, *SD* or *SE*, and the number of participants tested).
10. The study was written in English.

Studies meeting all 10 criteria were included in the meta-analysis. After applying the inclusion criteria to the retrieved reports, a total of 22 studies ($N = 2{,}202$) reporting a total of 139 posttest scores met the inclusion criteria and were included in the analysis. The studies consisted of 17 journal articles, 2 book chapters, 2 doctoral dissertations, and 1 master's thesis. Information about these studies is presented in Table 1.

Because published studies tend to report significant findings with larger ESs, the present study included both published and unpublished works. The advantage of this approach is that it minimizes publication bias.

Coding

Twenty-two studies were coded for study identification (i.e., author, publication year, and experiment number), activity type, test format, and dependent and moderator variables (see Appendix A for the overview of the coding scheme).

Activity Type. Independent variables were activity type and test format. Activities were coded as one of four types: (a) fill-in-the-blanks, (b) writing, (c) flashcards, and (d) word lists. Fill-in-the-blanks included activities in which participants encountered a text or sentences with some words omitted. Participants were either given lists of target words and their corresponding first language (L1) translations, or used a dictionary to look up the meanings of target words (e.g., Laufer, 2003), and then were instructed to write the appropriate target words in the blanks.

Writing refers to writing activities in which participants were instructed to write sentences or a composition using specific target words. Participants were provided with a list of target words and their corresponding L1 translations. Sometimes, participants were provided with a list of target words without their L1 translations and had to look up the meanings of the target words using a dictionary (Ansarin & Bayazidi, 2016).

In flashcard activities, participants would see target words and then be instructed to recall their meanings or see meanings and then try to recall target words. The flashcards could be either electronic or paper.

Word lists refer to activities in which participants were presented with lists of target words together with their meanings and were instructed to learn the target words. The difference between flashcards and word lists is that in the former, participants must retrieve information (L2 form or meaning) from memory, whereas in the latter both are presented together and so retrieval may not necessarily occur.

Test Format. The other independent variable was test format. Among the 22 included studies, 18 studies (81.8%) administered

TABLE 1
Basic Information About the 22 Included Studies

Study	Activity	Test Format	Test Timing[a]	L1–L2	Participant Place of Study	Proficiency
Al–Hadlaq (2003)	Fill-in-the-blank, writing	Meaning recall	Delayed posttest (5)	Arabic–English	University	Beyond basic
Ansarin & Bayazidi (2016)	Fill-in-the-blank, writing	Meaning recall	Immediate posttest	Mixed–English	University	Beyond basic
Bao (2015)	Writing	Meaning recall	Immediate posttest	Chinese–English	University	Beyond basic
Hulstijn & Laufer (2001): Experiment 1	Fill-in-the-blank, writing	Meaning recall	Immediate posttest, delayed posttest (7)	Dutch–English	University	Beyond basic
Hulstijn & Laufer (2001): Experiment 2	Fill-in-the-blank, writing	Meaning recall	Immediate posttest, delayed posttest (14)	Hebrew–English	University	Beyond basic
Javanbakht (2011)	Fill-in-the-blank, writing	Meaning recall	Immediate posttest, delayed posttest (7)	Persian–English	Secondary school	Basic
Keating (2008)	Fill-in-the-blank, writing	Meaning recall, form recall	Immediate posttest, delayed posttest (14)	English–Spanish	University	Basic
Keyvanfar & Badraghi (2011)	Fill-in-the-blank, writing	Meaning recall	Immediate posttest, delayed posttest (14)	Persian–English	Not reported	Beyond basic
Laufer (2003): Experiment 1	Writing	Meaning recall	Immediate posttest, delayed posttest (14)	Hebrew–English	University	Beyond basic

TABLE 1 (*Continued*)

Study	Activity	Test Format	Test Timing[a]	L1–L2	Participant Place of Study	Proficiency
Laufer (2003): Experiment 2	Writing	Meaning recall	Immediate posttest, delayed posttest (14)	Hebrew–English	University	Beyond basic
Laufer (2003): Experiment 3	Fill-in-the-blank, writing	Meaning recall	Immediate posttest, delayed posttest (14)	Arabic–English	Secondary school	Beyond basic
Mondria (2003)	List	Meaning recall	Immediate posttest, delayed posttest (14)	Dutch–French	Secondary school	Beyond basic
Mondria & Wiersma (2004)	List	Meaning recall, form recall	Immediate posttest, delayed posttest (13)	Dutch–English	Secondary school	Beyond basic
Nakata (2008)	List, flashcard	Form recall	Immediate posttest, delayed posttest (4)	Japanese–English	Secondary school	Beyond basic
Nakata (2016)	Flashcard	Meaning recall, form recall	Immediate posttest, delayed posttest (7)	English–Swahili	University	Basic
Pichette, de Serres, & Lafontaine (2012)	Writing	Form recall	Immediate posttest, delayed posttest (7)	French–English	University	Beyond basic
Rimrott (2010): Experiment 1	Flashcard	Form recall	Immediate posttest, delayed posttest (7)	English–German	University	Basic
Rimrott (2010): Experiment 2	Flashcard	Form recall	Immediate posttest, delayed posttest (7)	English–German	University	Basic

TABLE 1 (*Continued*)

Study	Activity	Test Format	Test Timing[a]	L1–L2	Participant Place of Study	Proficiency
Rott (2012)	Fill-in-the-blank, writing	Meaning recall, form recall	Immediate posttest, delayed posttest (14)	English–German	University	Beyond basic
Seibert (1927)	List	Form recall	Immediate posttest	English–French	University	Beyond basic
Tu (2004)	Fill-in-blank, writing	Meaning recall	Immediate posttest, delayed posttest (7)	Chinese–English	Secondary school	Beyond basic
Waring (1997)	Flashcard	Meaning recall, form recall	Immediate posttest, delayed posttest (7)	Japanese–English	University	Beyond basic
Webb (2005): Experiment 1	Writing, list	Meaning recall, form recall	Immediate posttest	Japanese–English	University	Beyond basic
Webb (2005): Experiment 2	Writing, list	Meaning recall, form recall	Immediate posttest	Japanese–English	University	Beyond basic
Webb (2007)	List	Meaning recall, form recall	Immediate posttest	Japanese–English	University	Beyond basic
Webb (2009a)	List	Meaning recall, form recall	Immediate posttest	Japanese–English	University	Beyond basic
Webb (2009b)	List	Meaning recall, form recall	Immediate posttest	Japanese–English	University	Beyond basic

Note. Information about activities or tests that are not related to the current meta-analysis was omitted.
[a] Number of days until the delayed posttest after the treatment is reported in parentheses.

meaning-recall tests and 13 studies (59.1%) administered form-recall tests. These test formats were included for analyses. Although several studies measured other aspects of vocabulary knowledge in addition to form–meaning connection (e.g., Bao, 2015; Coomber et al., 1986; Webb, 2005, 2009a), there was insufficient data for reliable analyses.

Of the 22 studies, 20 (90.9%) measured participants' learning gains immediately after engaging in activities, and 15 studies (68.2%) measured vocabulary retention in delayed posttests: Two studies (9.1%) measured gains 4–5 days later, seven studies (31.8%) measured learning 7 days later, and seven studies (31.8%) administered tests 12–14 days later.

Effect Size Calculations. We calculated ESs using the reported mean scores of immediate posttests and delayed posttests from studies in which students learned unknown target words. To evaluate the effectiveness of word-focused-activities from different perspectives, two types of ESs were calculated: (a) proportion of target words learned, and (b) number of target words learned per minute. To calculate the sample variances for each ES, the reported *SD* was transformed for each type of ES (see Appendix B for the calculation formulas).

Researchers tend to report multiple posttest scores from multiple activities, different types of tests, and/or different scoring methods. We calculated more than one ES per study when possible. Overall, a total of 139 reported posttest scores from 22 studies were used to calculate ESs included for analysis.

Moderator Variables. Twenty-two studies were further coded for moderator variables to investigate the effectiveness of the activities in relation to (a) participant characteristics, (b) activity characteristics, and (c) methodological features. Following Boulton and Cobb's (2017) approach, 16 studies were independently coded by two researchers who specialized in L2 vocabulary studies, and the number of disparities between the two researchers was calculated. The agreement rate was 98.5%. After all discrepancies were resolved through discussion, the second author carefully coded the rest of the six studies.

Participant Characteristics. Region where the research was conducted (i.e., Asia, Europe, Middle East, North America, Oceania) was coded. Participants in 21 of the 22 studies were learning English as a foreign language (Pichette et al., 2012, was the exception). This made it impossible to include foreign/second language as a variable.

L2 proficiency was also coded. Following Jeon and Yamashita's (2014) approach, proficiency was coded dichotomously—basic or beyond basic—to avoid inconsistency in judgement using different criteria (e.g., TOEFL, Vocabulary Levels Tests, teachers' intuitive judgements). When participants lacked any prior experience studying a target language (e.g., Nakata, 2016), or their proficiency levels were explicitly reported as "beginners" by researchers, the participants were coded as basic (e.g., Keating, 2008). The remaining participants were coded as beyond basic.

As a third participant variable, learner place of study was coded as either (a) secondary school or (b) university. The included studies did not deal with other types of institutions.

Activity Characteristics. Learning direction—receptive or productive learning—was coded as an activity-related variable. Learning direction was applied to the bidirectional activities word lists and flashcards that can be done both receptively and productively. The activities were coded as productive when participants were told to look at the L1 meanings and try to recall the L2 forms (L1 = > L2; e.g., Waring, 1997; Webb, 2009b). When told to learn the meanings of L2 words by recalling their respective L1 translations (L2 = > L1), the learning direction was coded as receptive.

Time on task (i.e., the number of target words presented divided by the time, in minutes, that students engaged in the activity) was coded as a second activity-related variable.

Methodological Features. Research design was coded as either (a) within-participants design, or (b) between-participants design. In studies using within-participants designs, participants engaged in more than one type of activity and studied more than one set of target words. Given that between-participants designs might be less demanding because participants engage in only one activity, studies using between-participants designs might have reported higher learning gains.

Scoring approach was coded as either (a) strict scoring, or (b) lenient scoring. Strict scoring refers to awarding points only for fully correct responses (e.g., Mondria & Wiersma, 2004). Lenient scoring refers to awarding points for partially correct responses (e.g., Laufer, 2003), or marking not only completely correct responses but also close approximations as correct (e.g., Webb, 2007).

As a third methodological feature, approach to determining preexisting knowledge of target words was coded. Eight studies (36.3%) measured participant knowledge using pretests or

questionnaires. Five studies (22.7%) tested knowledge with students of similar or higher proficiencies. Five studies (22.7%) used nonwords to ensure participants had no knowledge of target words. One study (4.5%) followed only researchers' reasoning based on the coursebooks used in the participants' school. Another study (4.5%) did not report how they ensured that participants did not know the target words. These last two studies were categorized as not checked.

The last methodological feature was the number of target words participants learned. This variable allowed us to examine whether the numbers of target words influenced the percentage of words learned.

Data Analysis

Dealing With Dependent Effect Sizes. Many studies included in the meta-analysis reported multiple posttest scores for different activities, test formats, and scoring methods. These nonindependent ESs violate the assumption of independent observation—that is, dependency or correlations among ESs can bias the variance estimation, potentially inflating Type I errors (e.g., Hox, 2010). As a solution to this issue, robust variance estimation (RVE; Hedges, Tipton, & Johnson, 2010) was adopted. With RVE, the inflation of Type I errors can be suppressed by adjusting estimated variances.

Analysis Procedure. To account for correlated ESs, RVE (Hedges et al., 2010) with correlated weights and small-sample adjustments (Tipton, 2015; Tipton & Pustejovsky, 2015) was used. The robumeta package (Fisher & Tipton, n.d.) and the clubSandwich package (Pustejovsky, 2018) were used in the R statistical environment (R Core Team, 2017). The metafor package (Viechtbauer, 2010) was also used to impute the sampling variance for each ES. Significance of the factors was examined using small-sample adjusted *t*-tests with the robumeta package and *F*-tests with the Wald_test function from the clubSandwich package.

To answer the first four RQs, which explore learning gains made through completing activities, two types of ESs (i.e., proportion of words learned, number of words learned per minute) were calculated and analyzed separately for meaning-recall and form-recall tests. Intercept-only models and no-intercept models including the activity type variable were fitted with the dataset subsequently to attain the aggregated ESs. In order to compare different activity types, post hoc tests with regression analysis changing reference levels (a.k.a. relevelling; de Vos et al., 2018) were administered.

To answer RQ5, meta-regression analyses were conducted to examine the influence of each moderator variable. In order to include the largest number of studies to increase statistical power, ESs of proportions of target words learned on immediate posttests were used as the dependent variable. Test format and activity type were inserted as covariates to appropriately assess the influence of moderator variables while avoiding potential biases (Lee, Warschauer, & Lee, 2018).[1] Only categories with more than three ESs were included for moderator analyses (Li, 2016). Because some moderator variables were not reported by all studies (e.g., time on task), these studies were excluded from moderator analyses.

Following de Vos et al. (2018), ESs were examined for Cook's distance larger than .85 to identify outliers. No ESs were identified as outliers. Potential publication bias was checked by conducting meta-regression analyses using published or unpublished as a moderator variable. Published comprised journal articles and book chapters. Unpublished comprised a master's thesis and doctoral dissertations. Because most of the studies included in the meta-analysis were published (19 studies, 86.4%), we analyzed ESs on meaning- and form-recall tests together in order to keep numbers of ESs as three or more for all analyses. Test format and activity type were used as covariates. Similarly, we conducted Egger's regression tests (Egger et al., 1997) to investigate the relationship between ESs and sampling variances while test format and activity type were used as covariates.

RESULTS

To answer RQ1 and RQ2, the proportion of target words learned and its relationship with two independent variables—activity type and test format—were analyzed. The estimate of the mean ES of proportion of the target words learned for all studies was 0.601, $SE = 0.045$, 95% CI [0.504, 0.697], $t(15) = 13.2$, $p < .001$, on meaning-recall tests, and 0.585, $SE = 0.057$, 95% CI [0.461, 0.710], $t(12) = 10.3$, $p < .001$, on form-recall tests. In order to calculate weighted mean ESs separately for each activity, a no-intercept model with activity variable was fitted. Table 2 presents the estimated ES on the two recall test formats. On the meaning-recall test, the effectiveness of the activities was in the following order: flashcards (77.0%), word lists (73.2%), writing (54.8%), and fill-in-the-blanks (43.1%). A Wald test with ESs on meaning-recall tests showed that activity type was

TABLE 2
Estimated Effect Size (ES) of Proportion of Target Words Learned on Immediate Posttests

	Meaning Recall				Form Recall			
Activity	*k*	*n*	Mean ES (SE)	CI	*k*	*n*	Mean ES (SE)	CI
Fill-in-the-blanks	8	9	0.431 (0.056)	[0.29, 0.56]	2	3	0.184 (0.042)	[–0.35, 0.72]
Writing	10	14	0.548 (0.018)	[0.43, 0.66]	4	6	0.368 (0.075)	[0.10, 0.62]
Word lists	5	11	0.732 (0.075)	[0.51, 0.94]	7	14	0.701 (0.051)	[0.57, 0.83]
Flashcards	2	6	0.770 (0.050)	[0.53, 1.00]	4	14	0.661 (0.048)	[0.50, 0.81]

Note. k = number of studies; n = number of ESs; SE = standard error; CI = 95% confidence interval adjusted with RVE. The total number of studies = 20. The total number of ESs = 77.

not significant, $F(3.46) = 7.42$, $p = .053$, but approached the traditional alpha level of .05. The reason statistical significance was not achieved in this study may be due to the number of studies with the variance being adjusted with RVE. Subsequent post hoc tests revealed that word lists ($p = .009$) and flashcards ($p = .022$) contributed to significantly larger gains than fill-in-the-blanks. The differences between fill-in-the-blanks and writing activities ($p = .078$) and between writing and word lists ($p = .078$), or between writing and flashcards ($p = .06$) approached significance, while there was no difference between word lists and flashcards ($p = .679$).

On the form-recall tests, the effectiveness of the activities was in the following order: word lists (70.1%), flashcards (66.1%), writing (36.8%), and fill-in-the-blanks (18.4%). A Wald test failed to detect significant differences among activity types, $F(1.91) = 12.7$, $p = .081$. Again, this was likely due to the numbers of studies while controlling the variance with RVE. The post hoc tests showed that flashcards and word lists led to significantly greater gains than fill-in-the-blanks ($p = .033$, $p = .043$, respectively) and writing ($p = .021$, $p = .020$). There were no significant differences between flashcards and word lists ($p = .578$) or between writing and fill-in-the-blanks ($p = .187$).

It should also be noted that there were considerable degrees of heterogeneity across ESs ($I^2 = 96.5$ and $I^2 = 96.6$ on meaning-recall and form-recall tests, respectively), showing that around 97% of the total variance is due to variance in the true effects of each word-focused activity and around 3% may be due to sampling variance. This indicates that learning gain differs considerably from study to study even when administering the same activity.

Fifteen studies administered posttests to measure vocabulary retention from 4 days to 2 weeks after the treatment. A total of 62 ESs from these studies were analyzed. The mean proportion of target words recalled on delayed posttests were 0.394, $SE = 0.06$, $t(11) = 6.55$, $p < .001$, 95% CI [0.261, 0.526], on meaning-recall tests and 0.251, $SE = 0.03$, $t(6.97) = 7.32$, $p < .001$, 95% CI [0.170, 0.332], on form-recall tests.

Table 3 presents the mean proportion of words recalled on delayed posttests separately for each activity. First, ESs on meaning-recall tests were analyzed. The proportion of words learned through completing each of the activities was as follows: flashcards (73.4%), word lists (47.9%), writing (31.9%), and fill-in-the-blanks (23.9%). A Wald test on ESs of meaning-recall tests detected significant differences among activities, $F(1.71) = 31.2$, $p = .046$. Post hoc tests showed that flashcards led to significantly greater retention than fill-in-the-blanks ($p = .007$), word lists ($p = .007$), and writing ($p = .012$). Word lists led to significantly greater gain than fill-in-the-blanks ($p = .040$). There was also a significant difference between writing and fill-in-the-blanks ($p = .040$).

The retention rate for each activity on form-recall tests was as follows: flashcards (32%), word lists (21.8%), fill-in-the-blanks (18.3%), and writing (18%). A Wald test did not detect significant differences among activities, $F(0.82) = 0.49$, $p = .761$. Subsequent post hoc tests also confirmed that there were no statistical differences among any pairs of activities (all ps > .10).

The influence of the number of days between treatment and delayed posttest (4 days to 2 weeks after the treatment) on test scores was examined. The results showed that test timing was not significantly related to the ES, either on meaning-recall ($p = .982$) or form-recall tests ($p = .650$).

Number of Target Words Learned per Minute

To answer RQ3 and RQ4, the number of target words learned per minute and its relationship

TABLE 3
Estimated Proportion of the Target Words Retained

	Meaning Recall				Form Recall			
Activity	*k*	*n*	Mean ES (SE)	CI	*k*	*n*	Mean ES (SE)	CI
Fill-in-the-blanks	8	12	0.239 (0.049)	[0.12, 0.35]	2	3	0.183 (0.056)	[–0.65, 1.01]
Writing	8	14	0.319 (0.047)	[0.20, 0.43]	3	14	0.180 (0.068)	[–0.15, 0.52]
Word list	2	5	0.479 (0.018)	[0.24, 0.71]	2	4	0.218 (0.024)	[–0.09, 0.53]
Flashcards	2	6	0.734 (0.012)	[0.58, 0.88]	4	4	0.320 (0.049)	[0.15, 0.48]

Note. ES = effect size; *k* = number of studies; *n* = number of ESs; CI = 95% confidence interval adjusted with RVE. The total number of studies = 15. The total number of ESs = 62.

TABLE 4
Estimated Effect Size (ES) of Number of Target Words Learned per Minute

	Meaning Recall				Form Recall			
Activity	*k*	*n*	Mean ES (SE)	CI	*k*	*n*	Mean ES (SE)	CI
Fill-in-the-blanks	6	7	0.176 (0.053)	[0.03, 0.31]	2	3	0.102 (0.038)	[–0.38, 0.58]
Writing	8	9	0.232 (0.050)	[0.11, 0.35]	3	4	0.256 (0.139)	[–0.37, 0.88]
Word lists	5	11	1.160 (0.177)	[0.65, 1.66]	7	14	1.278 (0.170)	[0.85, 1.70]
Flashcards	2	6	1.100 (0.029)	[0.73, 1.47]	4	14	0.775 (0.122)	[0.37, 1.17]

Note. *k* = number of studies; *n* = number of ESs; CI = 95% confidence interval adjusted with RVE. The total number of studies = 17. The total number of ESs = 68.

with activity type and test format were analyzed. Among 20 studies measuring immediate posttest scores, 17 studies reported the study time participants spent on the activity. A total of 68 ESs from those studies were analyzed. The average number of words learned per minute was 0.66, $SE = 0.14$, $t(13) = 4.49$, $p < .001$, 95% CI [0.34, 0.98], on meaning-recall tests and 0.91, $SE = 0.15$, $t(11) = 5.74$, $p < .001$, 95% CI [0.56, 1.26], on form-recall tests.

The aggregated number of words learned per minute for each activity is presented in Table 4. On meaning-recall tests, the efficacy of each activity was as follows: word lists (1.16), flashcards (1.10), writing (0.23), and fill-in-the-blanks (0.18). A Wald test showed that learning gain differed significantly among activities, $F(3.21) = 58.6$, $p = .003$. Subsequent post hoc tests revealed that both word lists and flashcards led to significantly larger gains than writing ($p = .002$, $p = .005$, respectively) and fill-in-the-blanks ($p < .001$, $p = .002$). There were no statistical differences between word lists and flashcards ($p = .768$) or between fill-in-the-blanks and writing ($p = .318$).

On form-recall tests, the efficacy of each activity was as follows: word lists (1.27), flashcards (0.77), writing (0.25), and fill-in-the-blanks (0.10). A Wald test showed no statistical difference among activities, $F(1.05) = 9.39$, $p = .224$. However, post hoc analyses found that word lists led to significantly greater gains than writing ($p = .034$). Also, the differences between word lists and fill-in-the-blanks, between word lists and flashcards, and between fill-in-the-blanks and flashcards did not reach but approached statistical significance ($p = .052$, $p = .052$, $p = .055$, respectively). In contrast, there was no significant difference between fill-in-the-blanks and writing ($p = .420$). There were also considerable heterogeneities in ESs even after activity type was accounted for ($I^2 = 98.1$ and $I^2 = 98.9$ on meaning-recall and form-recall tests, respectively).

Among 15 studies administering delayed posttests, 12 reported the study time participants spent on the activity. A total of 55 ESs from those studies were analyzed. The average number of words learned per minute on the delayed posttests was 0.32, $SE = 0.12$, $t(8.93) = 2.74$, $p = .023$, 95% CI [0.06, 0.58], on meaning-recall tests, and 0.28, $SE = 0.07$, $t(5.96) = 3.80$, $p = .009$, 95% CI [0.09, 0.46], on form-recall tests.

TABLE 5
Estimated Effect Size (ES) of Number of Target Words Learned per Minute on the Delayed Posttests

	Meaning Recall				Form Recall			
Activity	*k*	*n*	Mean ES (SE)	CI	*k*	*n*	Mean ES (SE)	CI
Fill-in-the-blanks	6	10	0.075 (0.021)	[0.019, 0.13]	2	3	0.104 (0.051)	[–0.542, 0.749]
Writing	6	10	0.079 (0.015)	[0.038, 0.119]	2	3	0.083 (0.001)	[0.069, 0.098]
Word lists	2	5	0.426 (0.082)	[–0.618, 1.471]	2	4	0.270 (0.091)	[–0.881, 1.42]
Flashcards	2	6	1.041 (0.015)	[0.851, 1.231]	4	14	0.396 (0.107)	[0.05, 0.742]

Note. k = number of studies; n = number of ESs; CI = 95% confidence interval adjusted with RVE. The total number of studies = 12. The total number of ESs = 55.

The aggregated number of words learned per minute for each activity on the delayed posttests is presented in Table 5. The number of words learned per minute on the meaning-recall tests for each activity was: flashcards, 1.04; word lists, 0.43; writing, 0.08; and fill-in-the-blanks, 0.08. A Wald test showed that learning gains significantly differed among activities, $F(1.19) = 333.04$, $p = .024$. Subsequent post hoc tests revealed that flashcard learning contributed to significantly more words learned per minute than fill-in-the-blanks ($p = .004$), writing ($p = .003$), and word lists ($p = .020$). Learning with word lists led to a greater number of words learned per minute than fill-in-the-blanks ($p = .072$) and writing ($p = .072$), with the differences approaching significance. There was no statistical difference between fill-in-the-blanks and writing ($p = .620$).

The number of words learned per minute on the form-recall test for each activity was: flashcards, 0.40; word lists, 0.27; fill-in-the-blanks, 0.10; and writing, 0.08. A Wald test showed no statistical difference among activities, $F(0.03) = 0.061$, $p = .979$. This was confirmed with subsequent post hoc tests, which did not find any statistical differences across activities (all ps $> .10$).

In answer to RQ5, Table 6 shows the results of the moderator analysis. Each estimated coefficient and its CI indicate differences from the respective reference levels in terms of the proportion of target words learned.

Characteristics of Participants

Three factors relating to participant characteristics were examined: (a) region, (b) L2 proficiency, and (c) place of study. Region was not significantly related to ES ($p = .506$), indicating that the effectiveness of word-focused activities may not differ across participants from different regions. L2 proficiency was also not significant ($p = .422$), suggesting that L2 students may benefit from word-focused activities regardless of their proficiency. Participant place of study was significant, $b = –.159$, 95% CI [–.292, –.025], $t(6.09) = –2.90$, $p = .027$. Secondary school students learned about 16% more target words than university students.

Characteristics of Activities

Two activity characteristics were analyzed: (a) learning direction, and (b) time on task. Learning direction was analyzed in bidirectional activities (i.e., activities such as word lists and flashcard learning that can be done receptively or productively). The results showed that the learning direction (receptive and productive) influenced learning gain differently on meaning-recall tests and form-recall tests. Learning direction was not significant on meaning-recall tests ($p = .864$). However, productive learning led to 22.4%, 95% CI [5.4%, 39.3%], greater gains than receptive learning on form-recall tests ($p = .019$). As for time on task, the number of minutes students spent per word was not significant ($p = .977$), indicating that longer study times do not necessarily lead to greater learning gains.

Methodological Features of Studies

Three methodological features of studies were analyzed: (a) research design, (b) scoring approach, and (c) determination of unfamiliarity of target words. Research design was not significant ($p = .525$), indicating that the ESs were not significantly different between studies using between-participant and within-participant designs. Scoring approach was also not significantly associated with the ES ($p = .919$), indicating that ESs were not significantly different between strict

TABLE 6
Moderator Analyses

Moderator Variables	*k*	*n*	Estimated Coefficients [CI]	Significance Test
Characteristics of participants				
Region				$F(1.07) = 2.73$, $p = .506$
Asia	8	30	–Reference level–	
Europe	2	8	.098 [–0.39, .0.59]	
Middle East	5	10	.162 [–.014, .338]	
North America	5	16	–.025 [–.169, .119]	
Oceania	1	11	–.077 [–.331, .176]	
Proficiency				$t(4.24) = -0.88$, $p = .422$
Basic	4	20	–Reference level–	
Beyond basic	16	57	–.078 [–.317, .161]	
Place of study				$t(6.09) = -2.90$, $p = .026^*$
Secondary school	5	15	–Reference level–	
University	15	60	–.159 [–.292, –.025]	
Characteristics of activities				
Learning direction				
Meaning recall				$t(4.20) = 0.032$, $p = .864$
Receptive learning	2	8	–Reference level–	
Productive learning	2	5	.011 [–.155, .177]	
Form recall				$t(5.36) = 11.05$, $p = .018^*$
Receptive learning	5	10	–Reference level–	
Productive learning	5	11	.224 [.054, .393]	
Time on task				$t(2.03) = 0.032$, $p = .977$
Minutes per word	17	68	.000 [–.091, .092]	
Methodological features				
Research design				$t(5.49) = -0.68$, $p = .525$
Between-participants design	15	56	–Reference level–	
Within-participants design	6	21	–.059 [–.277, .159]	
Scoring approach				$t(9.90) = 0.10$, $p = .919$
Strict scoring	9	27	–Reference level–	
Lenient scoring	18	50	.004 [–.100, .110]	
Method to check preexisting knowledge				$F(3.17) = 0.86$, $p = .543$
Nonwords	5	24	–Reference level–	
Pretests and questionnaires	8	32	.015 [–.193, .222]	
Other students	5	13	.170 [–.070, .410]	
Not checked	2	8	.069 [–.377, .516]	
Number of target words				
Target word number	20	77	.000 [–.091, .092]	$t(2.03) = 0.03$, $p = .977$

Note. *k* = number of studies; *n* = number of effect sizes; CI = 95% confidence interval adjusted with RVE. The total number of studies = 20. The total number of ESs = 77.
*$p < .05$.

and lenient approaches to scoring. Third, no significant differences were found among the different approaches to determining prior knowledge of target vocabulary ($p = .543$). However, it should be noted that studies checking participants' prior knowledge using other students (as opposed to actual participants) produced 17%, 95% CI [–7.0, 41.0], higher learning gains compared to studies using nonwords. Finally, the number of target words presented to participants was not significantly related to learning gain ($p = .977$).

Publication Bias

The results of meta-regression analyses using published or unpublished as a moderator variable did not reveal a publication bias for the

proportion of target words learned on the immediate posttests or number of words learned per minute on both immediate and delayed posttests ($ps > .10$). However, the mean ES of proportion of target words learned on delayed posttests was 16.7% higher for published studies compared to unpublished studies ($p = .022$). Egger's regression tests examining the relationship between ESs and sampling variances showed the same trend: There was no clear bias found ($p > .10$) except for the ESs as proportion of target words learned on the delayed posttests ($p = .044$).

DISCUSSION

In answer to RQ1, the findings indicate that intentional vocabulary-learning techniques contribute to relatively large learning gains on immediate posttests. The mean proportion of target words learned was 60.1% on meaning-recall tests and 58.5% on form-recall tests. This indicates that, on average, more than half of the target words were learned through completing word-focused activities. Although the results of the immediate posttests suggest that intentional vocabulary learning is quite effective, the results of the delayed posttests indicate that the long-term gains through intentional study are much smaller. The analyses revealed that 39.4% and 25.1% of target words were learned on meaning-recall and form-recall delayed posttests, respectively.

This finding is important because it shows that a single session of form-focused vocabulary learning on its own does not ensure learning. Instead, it should be viewed as the beginning of the word learning process. Initial gains in knowledge of form–meaning connection made through a word-focused activity might be relatively large. However, it is the subsequent encounters with partially known target words that are likely key to retaining and expanding upon that knowledge. There is a great deal to learn about each word, and the development of that knowledge is likely to be a gradual process that occurs through repeated encounters (Webb & Nation, 2017).

Incidental vocabulary-learning gains are typically viewed as being small in comparison to intentional learning gains (e.g., Laufer, 2003; Webb & Nation, 2017). However, there are several studies that have found delayed incidental learning gains that exceed those found for the activities examined in this meta-analysis (e.g., Cho & Krashen, 1994; Webb & Chang, 2015). It is important to note, however, that learning words in intentional or meaning-focused exercises rarely occurs in isolation; students may study a word in one activity, read about it in another, and study it further in more activities. These combined approaches to learning words will likely boost learning far beyond the gains found through completing a single activity.

RQ2 compared word learning through different intentional vocabulary-learning techniques. The mean proportions of words learned through completing each activity provides some indication of their relative efficacy. The mean proportions of words learned was in the following order on immediate meaning-recall tests: flashcards (77.0%), word lists (73.2%), writing (54.8%), and fill-in-the-blanks (43.1%). The analyses indicated that learning through word lists and flashcards contributed to greater gains than fill-in-the-blank activities. The mean proportions of words learned on the form-recall tests was word lists, 70.1%; flashcards, 66.1%; writing, 36.8%; and fill-in-the-blanks, 18.4%. The analyses indicated that learning with word lists and flashcards led to greater gains than learning with fill-in-the-blanks and writing activities.

The results of the meaning-recall delayed posttests again showed a great deal of variation among the relative efficacy of the activities. The order in the proportion of words learned was as follows: flashcards (73.4%), word lists (47.9%), writing (31.9%), and fill-in-the-blanks (23.9%). The post hoc analysis showed that learning with flashcards led to greater learning of the form–meaning connections of words than fill-in-the-blanks, writing, and word lists. Gains were also greater through word lists and writing than fill-in-the-blanks activities. There were also large differences among the activities in the proportions of words learned as indicated by the form-recall delayed posttests. The retention rate for each activity was: flashcards, 32%; word lists, 21.8%; fill-in-the-blanks, 18.3%; and writing, 18%. The post hoc analysis indicated that there was little difference among any of the activities on the form-recall delayed posttests.

The large variation in relative efficacy of the activities shows that all intentional vocabulary-learning tasks should not be considered equally effective for learning the form–meaning connections of words. The results indicated that flashcards are very effective. In contrast, proportional learning gains were far smaller for writing and fill-in-the-blanks. This highlights the value of learning with flashcards and also suggests that teachers and learners should carefully consider the activities that they choose for learning the form–meaning connections of words. It also indicates that there may be a great deal of value

in using frameworks such as Technique Feature Analysis (Nation & Webb, 2011) and TOPRA (Barcroft, 2015) that were designed to evaluate the efficacy of activities. There is a growing body of research indicating that Technique Feature Analysis can accurately predict task efficacy (e.g., Hu & Nassaji, 2016; Zou et al., 2018). Using frameworks may therefore guide teachers to use more effective activities.

It is also important to note that although the analyses indicated variation in the proportional word learning among the activities, these gains are specifically for knowledge of form–meaning connection because this has tended to be the aspect of vocabulary knowledge that is measured in studies of vocabulary learning. Moreover, transfer-appropriate processing theory suggests that the similarity between learning from flashcards and word lists and the form- and meaning-recall test formats commonly used in the research—as well as the difference between the fill-in-the-blank and writing conditions and these test formats—may at least partially account for the variation in gains among the activities. To accurately assess the relative contributions of activities, there is a clear need to measure vocabulary learning using other test formats (Webb, 2005). Form–meaning connection is only one of nine aspects in Nation's (2013) framework of vocabulary knowledge, and other aspects of knowledge might also be measured when evaluating vocabulary learning. The degree to which knowledge of collocation, grammatical functions, association, and word parts are learned in different activities has been examined in only a small number of word-focused studies. However, it is intuitively logical that contextualized activities such as composition writing and fill-in-the-blanks might make greater contributions to these aspects of knowledge than decontextualized activities such as flashcards and word lists.

In answer to RQ3, the analyses indicated that intentional vocabulary learning is efficient. On the immediate posttests, the average number of words learned per minute was 0.66 on meaning-recall tests, and 0.91 on form-recall tests. The greater number of words learned per minute on form-recall than meaning-recall tests was unexpected because research indicates that gains tend to be greater on meaning-recall tests (e.g., Laufer & Goldstein, 2004). This finding might have occurred because form-recall tests tend to be used less, as they are very demanding and consequently result in floor effects (e.g., Nation & Webb, 2011). Form-recall tests may therefore be used most often with activities that tend to produce relatively large gains. On the delayed posttests, the average number of words learned per minute was 0.32 and 0.28 on the meaning- and form-recall tests, respectively. Over 60 minutes, these figures translate to 19.2 and 16.8 words learned per hour on the meaning- and form-recall delayed posttests. These figures support Schmitt's (2000) and Nation's (2013) claims that intentional learning is efficient.

In answer to RQ4, the results revealed a great deal of variation among the activities in the average number of words learned per minute. On the immediate meaning-recall posttests, participants who learned with word lists had the highest rate of learning (1.16), followed by flashcards (1.10), writing (0.23), and fill-in-the-blanks (0.18). Post hoc analysis revealed that both word lists and flashcards led to a larger number of words learned per minute than writing and fill-in-the-blank activities. On the immediate form-recall posttests, the analyses revealed that the number of words learned per minute among the activities occurred in the same order as on the meaning-recall test: word lists (1.27), flashcards (0.77), writing (0.25), and fill-in-the-blanks (0.10). On the delayed meaning-recall test, the number of words learned per minute ranged from a high of 1.04 for flashcards to a low of 0.08 for both writing and fill-in-the-blanks. The differences in the number of words learned per minute when learning with flashcards was higher than for the other three activities. On the delayed form-recall test, the number of words learned per minute was in the order of flashcards (0.40), word lists (0.27), fill-in-the-blanks (0.10), and writing (0.08), with relatively little difference across the activities.

The variation in words learned per minute among the activities again reveals that we should be cautious about generalizing across intentional learning tasks. The results suggest that flashcards and word lists are extremely efficient in developing knowledge of form–meaning connection. The findings for the delayed tests indicated that flashcards would lead to gains in meaning and form recall of 62.4 and 24.0 words per hour, respectively. Word lists would lead to gains of 25.8 and 16.2 words per minute in meaning and form recall, respectively. In contrast, writing and fill-in-the-blanks are far less efficient in helping learners develop knowledge of form–meaning connection. Writing would contribute to learning 4.8 words per hour while fill-in-the-blanks would contribute to learning 4.8 to 6.0 words per hour. It should

be noted that although the gains for writing and fill-in-the-blanks are relatively small, they are still larger than those often found in studies of incidental learning through reading or viewing. For example, Horst, Cobb, and Meara (1998) found that participants learned 4.6 words through reading for 6 hours and Rodgers and Webb (2020) found that participants learned 6.4 words through viewing television for 7 hours. These figures translate to 0.013 and 0.015 words learned per minute, respectively.

In answer to RQ5, the results revealed that only two of the moderator variables (place of study, learning direction) were related to the proportion of words learned on immediate posttests, and that six variables (region, L2 proficiency, time on task, research design, scoring approach, and determination of prior knowledge of target words) were not. Surprisingly, the analyses indicated that secondary-school students made larger proportional gains in vocabulary knowledge than university students. This result might be explained by differences in the target words selected for studies with students at different institutional levels. It is intuitively logical that target words selected for learning by secondary students are higher in frequency and more concrete than those chosen for learning in studies with university students, which may make them easier to learn. Another possibility is that secondary-school students make greater use of rote-learning strategies than university students and that this has a positive effect on their gains in form–meaning connection (Dóczi, 2011). The finding that participants who learned in the productive direction in bidirectional activities had 16% higher gains than those who learned in the receptive direction was expected. Researchers have suggested that if there is only time to learn in one direction, it is better for students to retrieve L2 words rather than their meanings (e.g., Webb, 2009a, 2009b). However, these results should be interpreted with caution since only two studies accounted for the ESs on receptive recall tests for each direction. This limited number of included studies suggests that more studies directly examining the relationship between learning direction and testing direction are required to draw a more definitive conclusion. Similarly, two other moderator analyses (i.e., region and method to check preexisting knowledge) consisted of less than three studies in a category. Although these analyses are meaningful because they indicate the trends of the data, caution should be exercised when evaluating the certainty of the findings.

LIMITATIONS AND FUTURE DIRECTIONS

It is important to note several limitations of this meta-analysis. First, transfer-appropriate processing theory suggests that similarity between learning and testing conditions helps learners to demonstrate their knowledge (Morris et al., 1977). This similarity should have a positive effect for learners in the flashcard conditions as well as word list conditions in which participants retrieved form or meaning from memory. However, the differences between writing and fill-in-the-blanks and meaning-recall and form-recall tests may have had a negative effect on learning. Because the percentage learning gains occurred in an order that transfer-appropriate processing theory would predict, further research investigating how these activities contribute to learning with other types of measures would be useful.

A second limitation is that in most studies of intentional vocabulary learning, the number of retrievals of each target word is not clearly controlled. For example, studies of learning from word lists (e.g., Webb, 2009a, 2009b) involve participants learning for a set period of time rather than from a set number of retrievals. This makes it difficult to know precisely how repetition and the number of retrievals may have affected learning in different tasks. Because research tends to show that repetition affects learning (e.g., Uchihara, Webb, & Yanagisawa, 2019), it would be useful to more carefully examine this variable in future studies of intentional vocabulary learning. Moreover, it should also be noted that word-related variables such as number of letters, concreteness, and pronounceability—which are also known to affect learning (e.g., Laufer, 1997)—were not examined in this study. This was because there was insufficient data (learning gains according to word-related features) available to allow analyses of word-related variables as a moderator variable. It would be useful for future studies of word-focused instruction to include word-related factors as a variable to clarify how they affect the size of learning gains made through completing different activities.

Another limitation is that the present study only examined gains in knowledge of form–meaning connection. This was because few intentional learning studies measure gains in other aspects of word knowledge, a trend also observed in a recent meta-analytic review of incidental vocabulary-learning studies (Uchihara et al., 2019). Nation (2013) listed nine different types of vocabulary knowledge, of which

form–meaning connection is only one type. Decontextualized activities such as flashcards and word lists are focused on learning form–meaning connection, whereas writing and fill-in-the-blanks may also focus learners on several other aspects of vocabulary knowledge, such as collocation, grammatical functions, and word parts, in addition to form–meaning connection. We should therefore be cautious about dismissing the contributions of activities that involve learners attending to multiple aspects of vocabulary knowledge because they are less effective at developing knowledge of form–meaning connection.

A fourth limitation is that the results of the studies included in the meta-analysis represent a cross-sectional view of learning activities in isolation. This may not always be representative of intentional vocabulary learning. Words are often learned through the completion of a series of intentional and meaning-focused activities. Thus, the findings likely represent a fraction of the learning that may occur in the lexical development of each word. Longitudinal studies of learning through the completion of a series of ecologically valid activities may provide a more accurate assessment of the extent to which vocabulary that is taught in the classroom is learned.

The present study also revealed several areas in need of further investigation. First, there is a need for greater focus on different word-focused tasks. There are many different ways to deliberately learn words and the four types of activities examined in this study represent only a fraction of these. In particular, it would be useful to further investigate activities that are common to L2 learning coursebooks such as matching, true-or-false, multiple-choice, crossword, sentence completion, and classification activities. Second, greater detail on the L2 proficiency levels of participants is necessary to allow future meta-analyses to more accurately evaluate the relationship between intentional vocabulary learning and proficiency. In particular, it would be useful for future studies to include greater numbers of advanced participants and clearly report participant levels in relation to established benchmarks (e.g., Common European Framework of Reference for Languages or American Council on the Teaching of Foreign Languages Proficiency Guidelines). Third, further investigation of the effects of intentional vocabulary-learning activities on different aspects of vocabulary knowledge is warranted. Although in recent years there has been greater emphasis on using multiple tests to evaluate vocabulary learning (Nation & Webb, 2011), the norm is still to use these tests to assess gains in form–meaning connection. However, to provide a precise measurement of the effects of tasks, it is necessary to test the different aspects of word knowledge that may be gained through completing the task. Fourth, following Larson–Hall and Plonsky's (2015) guidelines for reporting on quantitative research findings, we strongly encourage researchers to make materials (e.g., activities, test formats, target words) and datasets publicly available. This would allow future meta-analyses to more accurately assess word-focused activities and the variables that moderate learning. Moreover, it would also enhance the transparency of the research, and allow more precise and robust estimations of individual participant data (see, e.g., Cooper & Patall, 2009).

CONCLUSION

Research findings on vocabulary-learning activities typically suggest that one activity is more effective than another. This can often lead to a discussion in which approaches are presented as two dichotomous options with one being the best choice. The aim of this meta-analysis was not to suggest that intentional learning is or is not the solution to L2 lexical development. L2 lexical development is a very long and complex process that may often involve learning words through a variety of intentional and meaning-focused activities. The present study has shown that the gains made through intentional vocabulary-learning activities tend to be relatively large on immediate posttests, but perhaps much smaller on delayed posttests than has been suggested in reviews of the literature on vocabulary learning (e.g., Nation, 2013; Schmitt, 2000, 2008). Moreover, the findings also indicate that activities contribute to learning in varying degrees. The results showed that both flashcards and word lists lead to relatively large gains in knowledge of form–meaning connection while writing and fill-in-the-blanks lead to relatively small gains. Indeed, the finding for flashcards and word lists are in line with the perception that intentional learning is effective and efficient. However, the results for the other two activities reveal that it would be a mistake to suggest that all types of activities are equally effective and efficient for gaining knowledge of the form–meaning connections of words. Taken together, the present study shows that teachers, learners, and researchers should not generalize learning efficacy across activities.

ACKNOWLEDGMENTS

We are most grateful to Akira Murakami, James E. Pustejovsky, Elizabeth Tipton, Wolfgang Viechtbauer, and Yo In'nami for their input on the statistical analysis. We would also like to thank the editors and reviewers for their useful comments.

NOTE

[1] When examining the moderator variable of learning direction, only activity type was used as covariate without test format. This is because the analyses were conducted separately with ESs on meaning-recall and form-recall tests. Furthermore, when examining the moderator variables for all activities, activity type was not used as covariate in order to examine whether learning direction influenced gains beyond the categorization of activity type.

Open Research Badges

This article has earned an Open Data badge. Open Data is available at https://www.iris-database.org.

REFERENCES

Al–Hadlaq, M. S. (2003). *Retention of words learned incidentally by Saudi EFL learners through working on vocabulary learning tasks constructed to activate varying depths of processing* (Unpublished doctoral dissertation). Ball State University, Muncie, IN.

Ansarin, A. A., & Bayazidi, A. (2016). Task type and incidental L2 vocabulary learning: Repetition versus task involvement load. *Southern African Linguistics and Applied Language Studies*, *34*, 135–146.

Azabdaftari, B., & Mozaheb, M. A. (2012). Comparing vocabulary learning of EFL learners by using two different strategies: Mobile learning vs. flashcards. *The Eurocall Review*, *20*, 47–59.

Bao, G. (2015). Task type effects on English as a foreign language learners' acquisition of receptive and productive vocabulary knowledge. *System*, *53*, 84–95.

Barcroft, J. (2004). Effects of sentence writing in second language lexical acquisition. *Second Language Research*, *20*, 303–334.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*, 35–56.

Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. Amsterdam/Philadelphia: John Benjamins.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*, 348–393.

Browne, C. M. (2003). *Vocabulary acquisition through reading, writing, and tasks: A comparison* (Unpublished doctoral dissertation). Temple University, Philadelphia, PA.

Cao, Z. (2013). The effects of tasks on the learning of lexical bundles by Chinese EFL learners. *Theory and Practice in Language Studies*, *3*, 957–962.

Cho, K., & Krashen, S. (1994). Acquisition of vocabulary from the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading*, *37*, 662–667.

Choi, M. L., & Ma, Q. (2015). Realising personalised vocabulary learning in the Hong Kong context via a personalised curriculum featuring "student-selected vocabulary." *Language and Education*, *29*, 62–78.

Coomber, J. E., Ramstad, D. A., & Sheets, D. R. (1986). Elaboration in vocabulary learning: A comparison of three rehearsal methods. *Research in the Teaching of English*, *20*, 281–293.

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

de Groot, A. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*, 463–506.

de Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning*, *68*, 906–941.

Dóczi, B. (2011). Comparing the vocabulary learning strategies of high school and university students: A pilot study. *WoPaLP*, *5*, 138–158.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*, 629–634.

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, *42*, 499–523.

Fisher, Z., & Tipton, E. (n.d.). *robumeta: An R-package for robust variance estimation in meta-analysis*. Accessed 4 March 2018 at https://www.researchgate.net/publication/273388328_robumeta_An_R-package_for_robust_variance_estimation_in_meta-analysis

Groot, P. J. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, *4*, 60–81.

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

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

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Hu, H. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement load hypothesis versus technique feature analysis. *System, 56*, 28–39.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *Modern Language Journal, 96*, 544–557.

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., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*, 539–558.

In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly, 44*, 169–184.

Javanbakht, Z. O. (2011). The impact of tasks on male Iranian elementary EFL learners' incidental vocabulary learning. *Language Education in Asia, 2*, 28–42.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*, 160–212.

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*, 365–386.

Keyvanfar, A., & Badraghi, A. H. (2011). Revisiting task-induced involvement load and vocabulary enhancement: Insights from the EFL setting of Iran. *Man & the Word/Žmogus Ir Žodis, 13*, 56–66.

Larson–Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Larson–Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*, 127–159.

Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 140–155). Cambridge: Cambridge University Press.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review, 59*, 567–587.

Laufer, B. (2005). Focus on form in second language vocabulary learning. In S. H. Foster–Cohen, M. d. P. G. Mayo, & J. Cenoz (Eds.), *EUROSLA yearbook* (Vol. 5, pp. 223–250). Philadelphia, PA: John Benjamins Publishing Company.

Laufer, B. (2006). Comparing focus on form and focus on formS in second-language vocabulary learning. *Canadian Modern Language Review, 63*, 149–166.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*, 399–436.

Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*, 1–26.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal, 28*, 89–108.

Lee, H., Warschauer, M., & Lee, J. H. (2018). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics, 40*, 721–753.

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition, 38*, 801–842.

Mondria, J.–A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings. *Studies in Second Language Acquisition, 25*, 473–499.

Mondria, J.–A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 79–100). Amsterdam/Philadelphia: John Benjamins.

Moonen, M., de Graaff, R., Westhoff, G., & Brekelmans, M. (2014). The multi-feature hypothesis: Connectionist guidelines for L2 task design. *Language Teaching Research, 18*, 474–496.

Morgan, J., & Rinvolucri, M. (2004). *Vocabulary*. Oxford: Oxford University Press.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL, 20*, 3–20.

Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning. *International Review of Applied Linguistics in Language Teaching, 54*, 257–289.

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? Effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition, 38*, 523–552.

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

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning from listening to L2 songs. *Studies in Second Language Acquisition, 41*, 745–768.

Pellicer–Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language, 22*, 31–55.

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through watching a single episode of L2 television. *Studies in Second Language Acquisition, 40*, 551–577.

Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics, 33*, 66–82.

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In *Advancing quantitative methods in second language research* (pp. 106–128). New York: Routledge.

Pustejovsky, J. (2018). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections (version 0.3.1). Accessed 9 April 2018 at 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.

Rimrott, A. (2010). *Computer-assisted vocabulary learning: Multimedia annotations, word concreteness and individualized instruction*. (Unpublished doctoral dissertation). Simon Fraser University: Burnaby, BC, Canada.

Rodgers, M. P. H., & Webb, S. (2020). Incidental vocabulary learning through watching television. *ITL—International Journal of Applied Linguistics, 171*, 191–220.

Rott, S. (2012). The effect of task-induced involvement on L2 vocabulary acquisition: An approximate replication of Hulstijn and Laufer (2001). In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 228–267). New York: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave McMillan.

Seibert, L. C. (1927). An experiment in learning French vocabulary. *Journal of Educational Psychology, 18*, 294–309.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*, 375–393.

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*, 604–634.

Tu, H.–F. (2004). *Effects of task-induced involvement on incidental vocabulary learning in a second language* (Unpublished master's thesis). National Tsing Hua University, Taiwan.

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

van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System, 41*, 609–624.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.

Waring, R. (1997). A study of receptive and productive learning from word cards. *Studies in Foreign Languages and Literature, 21*, 94–114.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*, 33–52.

Webb, S. (2007). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research, 11*, 63–81.

Webb, S. (2009a). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review, 65*, 441–470.

Webb, S. (2009b). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal, 40*, 360–376.

Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–239). New York: Routledge.

Webb, S., & Chang, A. C. S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition, 37*, 651–675.

Webb, S., & Nation, I.S.P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.

Won, M. (2008). *The effects of vocabulary instruction on English language learners* (Unpublished doctoral dissertation). Texas Tech University, Lubbock, TX.

Zou, D., Wang, F. L., Kwan, R., & Xie, H. (2018). Investigating the effectiveness of vocabulary learning tasks from the perspective of the technique feature analysis: The effects of pictorial annotations. In S. K. S. Cheung, J. Lam, K. C. Li, O. Au, W. W. K. Ma, & W. S. Ho (Eds.), *Technology in education: Innovative solutions and practices* (pp. 3–15). Singapore: Springer Nature Singapore.

APPENDIX A

Coding Scheme

Variables	Values			
Identification				
Author				
Title				
Year of publication				
Type of publication	Journal article	Master's thesis	Dissertation	Book/ book chapter
Independent variables				
Activity type	Fill-in-the-blanks	Writing	Flashcards	Word lists
Test format	Meaning recall	Form recall		
Test timing	Immediate	Delayed		
Test date				
Moderator variables				
Participant characteristics				
Region	Asia	Middle East	North America	Oceania
Learner place of study	Secondary school	University		
Proficiency	Basic	Beyond basic		
Activity characteristics				
Learning direction	Receptive	Productive		
Time on task (minutes)				
Methodological features				
Research design	Between-participants	Within-participants		
Scoring approach	Strict scoring	Lenient scoring		
Approach to determining preexisting knowledge of target words	Pretests or questionnaires	Other students	Nonwords	Not checked
Number of target words learned				
Dependent variables				
Mean of posttest scores				
SD for posttest score				
Maximum posttest score				
Number of participants tested				
Number of target words				
Activity time (min)				

Note. Variables without labelled values are continuous, noncategorical, or open-ended.

APPENDIX B

Calculation Formulas for Effect Sizes and Standard Deviations

Proportion of target words learned:

$$\text{ES (proportion)} = \frac{\text{Mean of posttest scores}}{\text{Maximum posttest score}}$$

$$\text{SD (proportion)} = \frac{\text{SD for posttest score}}{\text{Maximum posttest score}}$$

Number of target words learned per minute:

$$\text{ES (number of target words learned per minute)} = \frac{\text{ES (proportion)} \times \text{number of target words}}{\text{Total activity time (minutes)}}$$

$$\text{SD (number of target words learned per minute)} = \frac{\text{SD (proportion)} \times \text{number of target words}}{\text{Total activity time (minutes)}}$$

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.