Studies in Second Language Acquisition, **44** (2022), 1279–1308
doi:10.1017/S0272263121000577

Research Article

INVOLVEMENT LOAD HYPOTHESIS PLUS

CREATING AN IMPROVED PREDICTIVE MODEL OF INCIDENTAL VOCABULARY LEARNING

Akifumi Yanagisawa

Tokyo University of Science

Stuart Webb

University of Western Ontario

Abstract

The present meta-analysis aimed to improve on Involvement Load Hypothesis (ILH) by incorporating it into a broader framework that predicts incidental vocabulary learning. Studies testing the ILH were systematically collected and 42 studies meeting our inclusion criteria were analyzed. The model-selection approach was used to determine the optimal statistical model (i.e., a set of predictor variables) that best predicts learning gains. Following previous findings, we investigated whether the prediction of the ILH improved by (a) examining the influence of each level of individual ILH components (need, search, and evaluation), (b) adopting optimal operationalization of the ILH components and test format grouping, and (c) including other empirically motivated variables. Results showed that the resulting models explained a greater variance in learning gains. Based on the models, we created incidental vocabulary learning formulas. Using these formulas, one can calculate the effectiveness index of activities to predict their relative effectiveness more accurately on incidental vocabulary learning.

The experiment in this article earned an Open Data badge for transparent practices. The materials are available at https://osf.io/4g5wy/

This article is based on part of the first author's doctoral dissertation submitted to the University of Western Ontario in 2020. We would like to thank Batia Laufer, Jan Hulstijn, Frank Boers, Donna Goldhawk, Perry Klein, Debra Jared, and Ruslan Suvorov for their helpful feedback on the current research project. We also thank Akira Murakami and James E. Pustejovsky for their input on statistical analysis. We are grateful to Takumi Uchihara, Tomlin Gagen, and Su Kyung Kim for their support in the process of data collection. Our gratitude also goes to the following researchers for providing information for the current meta-analysis: Jeanine Treffers-Daller, Chizuru Mori, Karim Jahangiri, Mandana Hazrat, Ali Jahangard, Mouhammad Reza Sarbazi, Mark Feng Teng, and Sasan Baleghizadeh.

*This article has been updated since its original publication. See https://doi.org/10.1017/S027226312200033X.

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

INTRODUCTION

Laufer and Hulstijn's (2001) Involvement Load Hypothesis (ILH) was designed to predict the effectiveness of instructional activities[1] on incidental vocabulary learning. The ILH posits that retention of L2 unknown words is contingent upon the *involvement load* (IL) of an activity. IL is determined by one motivational factor (*need*) and two cognitive factors (*search* and *evaluation*). The ILH predicts that the effect of an activity increases as the degree to which these factors in the learning condition increase. The ILH has frequently been referred to in order to provide pedagogical suggestions on how to select and design effective activities for learning new words (e.g., Barclay & Schmitt, 2019; Coxhead, 2018; Newton, 2020; Webb & Nation, 2017).

Many studies have tested how accurately the ILH predicts the relative effectiveness of activities. The majority of studies provided general support for the ILH by finding that students tended to learn more words from activities with higher ILs compared to activities with lower ILs (e.g., Eckerth & Tavakoli, 2012; Hulstijn & Laufer, 2001; Kim, 2008; Kolaiti & Raikou, 2017; Laufer, 2003). However, several studies also revealed that the predictions of the ILH were not always accurate (e.g., Bao, 2015; Folse, 2006; Keating, 2008; Rott, 2012; Zou, 2017). These studies argued that the individual components (need, search, and evaluation) might contribute to learning differently (e.g., Kim, 2008; Laufer & Hulstijn, 2001) and other factors (e.g., frequency, mode of activity, and test format) should also be included (e.g., Folse, 2006). To evaluate the predictive ability of the ILH, Yanagisawa and Webb (2021) adopted a meta-analytic approach to statistically summarize studies that tested the ILH's prediction. The results largely supported the ILH by finding that there was a clear pattern showing that learning gains increased as the IL of activities increased. However, the results also showed that the ILH explained a limited amount of variance in learning gains. Furthermore, each component of the ILH (need, search, and evaluation) contributed to learning at varying degrees. The results also showed that other factors (e.g., frequency and test format) influenced incidental vocabulary learning in addition to the IL of tasks. These findings raised the possibility that the predictive ability of the ILH could be enhanced by evaluating the relative influence of each ILH component and by considering other influential factors. Including other factors would provide a more comprehensive framework that could predict vocabulary learning. Therefore, the present study aims to determine whether such a framework would enhance the accuracy in predicting incidental vocabulary learning. Candidate influential factors as well as the ILH components are analyzed by using a model selection approach to obtain a statistical model including a combination of predictor variables that meaningfully contribute to the prediction of learning gains. Based on this resulting model, we aim to create formulas to calculate the effectiveness index of activities. The effectiveness index predicts the relative effectiveness of activities for incidental vocabulary learning. Future individual studies can assess the predictive power of the proposed formulas by testing them as falsifiable hypotheses in the same manner as the original ILH.

BACKGROUND

The ILH claims that retention of unknown L2 words is determined by the degree to which three factors in a learning condition are present: need, search, and evaluation. Activities

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

involving higher degrees of these factors are predicted to elicit greater vocabulary learning than activities involving lower degrees. *Need* is the motivational factor relating to whether a word is needed to complete the activity. Need has three levels: (a) absent when the unknown word is not needed to complete the activity (0 points), (b) moderate when an external entity (e.g., activity or teacher) asks students to understand or use the word (1 point), and (c) strong when the need for the word is derived by the learners, for example, wanting to know or use the words (2 points). For example, need is moderate when an activity requires a student to use an unknown word in a sentence. In contrast, need is strong when a student consults with a dictionary to look up an unknown word because they want to use the word in speech or in writing.

Search is a cognitive factor regarding the act of searching for a word. Search has two levels: presence or absence. Search is present when a student is required to search for L2 form or meaning using external resources (e.g., dictionaries, peers, or teachers) (1 point). Search is absent when L2 form and meaning are provided together in a task (0 points). One example of an activity that includes search is reading a text while looking up unknown words using a dictionary. In contrast, search is absent if students are provided with glosses near unknown words so there is no need to search for their forms or meanings.

Evaluation is another cognitive factor involving the comparison of a word's L2 form or meaning with other words or meanings to select the most suitable one for a specific context. Evaluation has three levels: absent, moderate, and strong. Evaluation is absent when there is no clear need to determine which word or meaning of a word to use (0 points). It is moderate when context is provided (1 point). One activity that includes moderate evaluation is fill-in-the-blanks, where students select the most suitable words for the blanks in a text while being provided with several options. Evaluation is strong when students have to use a word in an original context. One example that includes strong evaluation is a sentence production activity (2 points).

The IL of an activity is indicated by the sum of the scores of the three components (Laufer & Hulstijn, 2001, p. 16). For instance, a reading activity, where students read sentences with glosses of target words and answer comprehension questions that require students to understand the words, has an IL of 1 (moderate need = 1 point, no search = 0 points, and no evaluation = 0 points). In contrast, a composition writing activity, where students have to use all target words in a composition with a list of target words and their meanings provided, has an IL of 3 (moderate need = 1, no search = 0, and strong evaluation = 2). Because the composition writing activity scores higher than the reading activity, the ILH predicts that the former would lead to less learning than the latter.

The ILH has two stipulations: Activities must involve incidental learning rather than deliberate learning, and other factors must be equal. The ILH predicts incidental vocabulary learning but not intentional vocabulary learning. Here, incidental learning is defined as learning that occurs while engaging in activities without the clear intention of committing target words to memory. In intentional learning situations in which students are forewarned of an upcoming vocabulary test, it may be challenging to predict the degree to which words might be learned because students may spend most of their time trying to remember the target words instead of appropriately pursuing the goal of the activity (e.g., reading for comprehension). Moreover, Laufer and Hulstijn (2001, p. 11) argue that in intentional learning, each student may use different strategies to remember

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

target words and learning gains may be reflected by the strategy that was used instead of the learning activity in which they engaged.

The ILH claims that when *other factors are equal*, words that are processed with a higher involvement load will be retained better than words that are processed with a lower involvement load. This means that when factors such as frequency and mode of input (written or spoken) are different across tasks, learning gains might not be as the ILH predicts. This stipulation is important because it clearly states the realm in which the ILH is designed to make reliable predictions of vocabulary learning. However, it may also be useful to consider whether the addition of other factors might increase the prediction of how well words will be learned. Classroom learning environments tend to include varying factors in addition to the IL of activities. Therefore, investigating a greater number of factors may also enable predictions in a wider variety of contexts.

EARLIER STUDIES TESTING THE PREDICTION OF THE ILH

Many studies have examined whether the ILH accurately predicts the relative effects of activities on vocabulary learning, directly (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2008; Rott, 2012) and indirectly (e.g., Folse, 2006; Lee & Hirsh, 2012). The studies have produced mixed results. Several studies have found that the relative effectiveness of activities was exactly as the ILH predicted; activities with a higher IL led to greater learning and activities with the same IL led to similar learning gains (e.g., Eckerth & Tavakoli, 2012; Kim, 2008; Tang & Treffers-Daller, 2016). For example, Kim (2008) examined the prediction of the ILH with L2 English learners in two different proficiency groups. She found that regardless of the proficiency, the activities with higher ILs led to greater learning than the activities with lower ILs, and activities with the same IL led to similar learning gains. Eckerth and Tavakoli (2012) examined the effects of IL and frequency. They examined three activities with varying ILs where students encountered target words at different frequencies, one or five. Their results supported the ILH by finding that both IL and frequency influenced learning and that the relative effectiveness of activities was in line with the prediction of the ILH. Support was also provided by Huang, Willson, and Eslami's (2012) meta-analysis of 12 studies comparing learning from output activities (e.g., gap-filling and writing) to input activities (i.e., reading). They found that output activities with higher ILs yielded greater learning gains than output activities with lower ILs, corroborating the prediction of the ILH.

In contrast, many studies yielded findings that were not entirely in line with the ILH's prediction. Several studies found that activities with higher ILs did not outperform activities with lower ILs (e.g., Martínez-Fernández, 2008; Yang et al., 2017), or activities with the same IL led to significantly different learning gains (e.g., Zou, 2017). Moreover, in some studies, activities with lower ILs outperformed activities with higher ILs (e.g., Bao, 2015; Wang et al., 2014). It is important to note that contrasting results have also occurred when recruiting multiple samples of participants or measuring learning gains with multiple test formats and/or different test timings. For example, Hulstijn and Laufer (2001) found that although the relative effectiveness of activities was as the ILH predicted in one experiment with English learners in Israel, another experiment with English learners in the Netherlands found that the prediction was only partially accurate. Rott (2012) measured learning with two test formats: receptive recall (L2 to L1 translation) and productive recall (L1 to L2 translation) tests. While the results of the productive test immediately administered after learning produced full support for the ILH prediction, those of the receptive test only partially supported the ILH.

One way to untangle the inconsistency in findings is to conduct a meta-analysis. By statistically summarizing the results of earlier studies, a meta-analysis can provide a more summative and reliable overview of the findings. The systematic procedure of meta-analysis enables a comprehensive literature search to provide a more objective summary of findings than a typical literature review (In'nami et al., 2020). Yanagisawa and Webb (2021) meta-analyzed earlier studies that tested the prediction of the ILH. They analyzed the 42 studies that met their criteria to determine the overall extent to which the ILH predicts incidental vocabulary learning gains (i.e., the proportion of unknown words learned). The results provided general support for the ILH by finding a clear correlation between ILs and learning gains, illustrating that learning increased as the IL of activities increased. However, the results also showed that the ILH explained a limited amount of the variance in learning gains. The variance explained at the within-study level—reflecting the differences in posttest scores within the same study—was 29.1% on immediate posttests and 26.5% on delayed posttests. Similarly, the total variance explained—reflecting the overall differences in posttest scores across studies—was 15.4% on immediate posttests and 5.5% on delayed posttests. These figures suggest that learning gains are also affected by factors other than those in the ILH. The meta-analysis also revealed that the individual components of the ILH (need, search, and evaluation) contributed to learning at varying degrees. Evaluation was found to contribute to the greatest amount of learning, followed by need. Search, however, was not found to contribute to learning. These findings are in line with Laufer and Hulstijn's (2001) suggestion for further research to compare tasks with the same number of components, but with a different distribution of the components involved, because not all three factors may be equally important for vocabulary learning.

POTENTIAL APPROACHES TO ENHANCING THE PREDICTION OF INCIDENTAL VOCABULARY LEARNING

Results of earlier studies testing the prediction of the ILH suggest potential approaches to enhancing the prediction of incidental vocabulary learning. The accuracy of the ILH's prediction can be enhanced by (a) evaluating the degree of influence of each ILH component and (b) revising the evaluation component. Furthermore, other factors can be added to the ILH to form a more comprehensive predictive model.

First, it might be possible to enhance the prediction of the ILH by assessing the degree of influence of each ILH component. The ILH postulates that some of the components contribute to learning to the same degree. Specifically, moderate need, moderate evaluation, and present search (as search is either present or absent) are all awarded 1 point and within the ILH are thus assumed to contribute to learning to the same degree. The same goes for strong need and strong evaluation, which are both awarded 2 points and thus assumed to have the same degree of influence. However, it may be possible that the individual components contribute to learning to different degrees. As stated earlier, Laufer and Hulstijn (2001) mentioned this possibility and recommended further investigation of the influence of each component. Several studies have also indicated that the

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

components might carry different weights. Kim (2008) argued that strong evaluation might contribute to learning to the greatest extent, while Tang and Treffers-Daller (2016) found that search might contribute less than need and evaluation. Yanagisawa and Webb's (2021) meta-analysis of the ILH captured this trend revealing that evaluation had the most substantial influence, followed by need, while search was not found to have influence on learning. It is also important to note that the ILH assumes that strong need and strong evaluation have a larger impact on learning than moderate need and moderate evaluation (2 points are awarded for both strong need and strong evaluation, while 1 point is awarded for moderate need and moderate evaluation). Examining the exact magnitude of each component may enhance the prediction of incidental vocabulary learning.

Second, revising the evaluation component might enhance the prediction. Zou (2017) examined vocabulary learning from three activities while manipulating evaluation: fill-in-the-blanks (moderate evaluation), sentence writing (strong evaluation), and composition writing (strong evaluation). The results showed that composition writing led to greater vocabulary learning than sentence writing even though the ILs of these activities were the same. Based on this finding and an analysis of interview and think-aloud data, Zou argued that evaluation might better be divided into four levels: no evaluation, moderate evaluation, strong evaluation (sentence level), and very strong evaluation (composition level). In contrast, Kim (2008) compared sentence writing and composition writing and found similar degrees of learning gains. It would be useful to use meta-analysis to examine the results of more studies testing the ILH to determine whether dividing evaluation into four levels increases ILH's prediction accuracy.

Third and lastly, adding other factors to the ILH might also enhance its prediction. Among many factors that potentially influence incidental vocabulary learning, five factors have been widely discussed and examined in the context of the ILH: frequency, mode of activity, test format, test day, and the number of target words.

Frequency

Several studies examined the prediction of the ILH while manipulating the frequency of encounters or use of target words (e.g., Eckerth & Tavakoli, 2012; Folse, 2006; Lee & Hirsh, 2012). Folse (2006) found that an activity with a lower IL but repetition of target items contributed to greater vocabulary learning than an activity with a higher IL and no repetition of target items. A similar finding was reported by Lee and Hirsh (2012), who argued that the number of word retrievals may be more important than the IL of an activity. Because studies sometimes tested the prediction of the ILH with varying frequencies of encounters and use of target items (e.g., Ansarin & Bayazidi, 2016, 3 times; Beal, 2007, 2 times; Martínez-Fernández, 2008, 4 times), a meta-analysis might be able to tease apart the effect of frequency from that of other factors to determine whether its inclusion in the suggested predictive model might enhance the prediction of learning gains.

Mode of activity

Although the majority of the ILH studies examined activities that involve reading and writing (e.g., reading, fill-in-the-blanks, and writing), several studies also included activities that involve listening and speaking (e.g., Jahangard, 2013, listening activities; Hazrat, 2015, speaking activities, and Karalik & Merç, 2016, retelling activities), or activities where students were provided with language input in both written and spoken modes (Snoder, 2017). For example, Hazrat (2015) compared oral sentence generation to sentence writing. The results showed that although both activities had the same IL, sentence writing led to greater word learning than oral sentence generation. There are few studies that have explicitly compared incidental vocabulary learning from spoken and written input. However, two studies have found that incidental vocabulary learning gains are larger through reading than listening (Brown et al., 2008; Vidal, 2011), while one study (Feng & Webb, 2020) found no difference between the gains made through these two modes. Thus, it may be hypothesized that learning gains from spoken activities produce lower learning gains than written activities.

Alternatively, there is also reason to believe that speaking and listening activities might lead to greater word learning than reading and writing activities. Two cognitive schemes, *Multimedia Learning Theory* (Mayer, 2009) and *Dual Coding Theory* (Sadoski, 2005; Sadoski & Paivio, 2013), suggest that processing information in both visual and verbal channels leads to better retention of target items than processing in either channel alone. In activities that incorporate speaking and listening (e.g., Hazrat, 2015; Jahangard, 2013; Karalik & Merç, 2016), students were often provided with the target words both in written and spoken forms (e.g., the provision of a glossary). As Multimedia Learning Theory and Dual Coding Theory would suggest, such spoken activities including both written and spoken modes might contribute to greater learning gains than written activities.

Test format

Because the sensitivity of tests greatly influences learning gains (e.g., Webb, 2007), accounting for how vocabulary knowledge was measured might enhance the prediction of learning. Meta-analyses tend to group different test formats to obtain the overall mean of learning gains for different aspects of vocabulary knowledge. For example, de Vos et al. (2018) grouped test formats into two groups: (a) recognition (multiple-choice questions) and (b) recall (meaning and form cued recall tests). Yanagisawa, Webb, and Uchihara (2020) added an *other tests* category to further distinguish tests focusing on form-meaning connection (i.e., recognition and recall) from tests that may tap into knowledge of other aspects of vocabulary knowledge (i.e., VKS and gap-filling tests). Studies testing the ILH have also measured vocabulary learning using several different test formats. Tests in these studies could be placed in four groups: receptive recall (e.g., Hulstijn & Laufer, 2001; Rott, 2012), productive recall (e.g., Hazrat, 2015; Rott, 2012), recognition (e.g., Martínez-Fernández, 2008), and other test formats (e.g., Bao, 2015; Kim, 2008), or each test format could be examined separately. Given that grouping test formats that have different sensitivities to learning may ambiguate learning gains and worsen the prediction, it is important to identify the optimal grouping of test formats. Because the current study examines ILH studies' reported learning gains measured with a variety of test formats, accounting for the effect of test format may increase the precision of the estimated effects of other variables (e.g., ILH components, mode, and frequency) on learning.

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

Test day

Research measuring learning gains at different timings tends to show that gains decrease as the number of days between learning and testing increase (e.g., Keating, 2008; Rott, 2012). This suggests that the time of testing may affect the accuracy of the ILH prediction. Therefore, it may be useful to examine the general trend of how learned words were forgotten by statistically summarizing the results of ILH studies. Moreover, including test day (the number of days between learning and testing) in the statistical model may enhance the accuracy of the prediction.

Number of target words

The number of target words in studies examining the ILH has varied (e.g., Folse, 2006, 5 words; Hulstijn and Laufer, 2001, 10 words; Bao, 2015, 18 words). It may be reasonable to assume that when students encounter or have to use more words in an activity, the time they have to learn each word decreases. Research suggests that the amount of attention paid to words during incidental learning activities affects learning; words that receive greater attention are more likely to be learned than those that receive less attention (e.g., Godfroid et al., 2013; Pellicer-Sánchez, 2016). There is yet insufficient data to incorporate the amount of attention paid to words as a factor into a meta-analysis of the ILH. However, it is possible to determine whether the inclusion of the number of target words as a factor enhances the accuracy of the prediction of learning.

Other factors have also been reported to influence incidental vocabulary learning (e.g., time on task, L2 proficiency, working memory, and the features of lexical items). Unfortunately, little data has been provided about these variables in studies testing the ILH, and to examine the effect of a variable by meta-regression analysis (especially with a model selection approach used by the current study), the variable has to be reported in all studies. The present study investigated frequency, mode of activity, test format, test day, and number of target words as additional factors that might add to the ILH prediction because data for these variables has been widely reported. The need for increased reporting of other factors will be further discussed in the "Limitations and Future Directions" section of this article.

THE CURRENT STUDY

Research has indicated that it would be useful to try to improve upon Laufer and Hulstijn's (2001) ILH framework. Yanagisawa and Webb (2021) found that although a clear correlation between learning and IL was found, the ILH explained a limited variance in learning gains. One way in which the ILH might be improved is through weighting the ILH components (e.g., Kim, 2008; Laufer & Hulstijn, 2001). A second way to enhance the predictive power of the ILH may be to distinguish between different types of evaluation (Zou, 2017). In addition, including other empirically motivated factors (e.g., frequency, mode, test format, and test day) might increase the explained variance of vocabulary learning (Folse, 2006; Hazrat, 2015; Rott, 2012).

The present study aims to expand the ILH to provide a more comprehensive framework that predicts incidental vocabulary learning. Through meta-analyzing studies that

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

examined incidental vocabulary learning gains while strictly controlling the ILs of tasks, we seek to identify the optimal statistical model that best predicts learning gains. Based on the resulting model indicating the effect of each predictor variable, we created incidental vocabulary learning (IVL) formulas. Future studies can test the prediction of the IVL formulas in the same manner as the original ILH.

This study was guided by the following research question:

1. What is the best combination of predictive variables for incidental vocabulary learning within studies investigating the effect of involvement load?

METHOD

DESIGN

To statistically analyze the results of earlier studies that examined the effect of IL on vocabulary learning, we adopted a meta-analytic approach. Following common practice in meta-analysis in applied linguistics (e.g., Plonsky & Oswald, 2015), we first conducted a literature search to identify studies that tested the prediction of the ILH where L2 students learn vocabulary incidentally. Second, the identified studies were filtered to exclusively include the studies that met our criteria and were appropriately analyzable with meta-regression. Third, studies were coded for their dependent variable (i.e., the reported learning gains) and predictor variables (e.g., ILH components and other factors that potentially influence vocabulary learning). Fourth, the reported learning gains were analyzed using a three-level meta-regression model (Cheung, 2014) with a model selection approach. The analysis procedure includes (a) identifying the best operationalization of the ILH, (b) identifying the best grouping of test formats, and (c) determining the optimal combination of variables that best predicts learning gains. This process enabled us to identify factors that meaningfully contribute to the prediction of learning gains. The resulting statistical model includes all of the identified meaningful predictors to statistically control for the influence of each predictor and increase the precision of the estimation of each predictor's effect. Lastly, based on the resulting model, we created two formulas to calculate the effectiveness index of activities similar to the original ILH and test its prediction as a falsifiable hypothesis.

DATA COLLECTION

Literature search

To comprehensively include studies that examined the effect of IL on incidental vocabulary learning, we followed previously suggested guidelines (In'nami & Koizumu, 2010; Plonsky & Oswald, 2015) and searched the following databases: Educational Resources Information Centre (ERIC), PsycINFO, Linguistics and Language Behavior Abstract (LLBA), ProQuest Global Dissertations, Google Scholar, and VARGA (at Paul Meara's website: http://www.lognostics.co.uk/varga). Unpublished research reports such as doctoral dissertations, master's theses, and book chapters were also included (Oswald and

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

Plonsky, 2010). Research reports published from 2001 to April 2019 were found using different combinations of keywords such as involvement load hypothesis, task-induced involvement, involvement load, word/vocabulary, learning/acquisition/retention, and task. Through the electronic database search, a total of 963 reports were identified. Furthermore, we conducted a forward citation search to retrieve studies citing Laufer and Hulstijn (2001) and including the keywords in their titles by using Google Scholar to search for the studies that examined vocabulary learning and potentially discussed the ILH. Through this forward citation search, 327 more reports were found. Consequently, a total of 1290 reports were identified.

Inclusion and exclusion criteria

The identified research reports were screened using the following six selection criteria to determine which studies to include.

1. Studies looking at vocabulary learning from incidental learning conditions were included. Following Hulstijn's (2001) and Laufer and Hulstijn's (2001) definition of incidental vocabulary learning, studies were included when participants were not forewarned about upcoming vocabulary tests before the treatment and participants were not told to commit target words to memory. We excluded studies where participants were told about posttests (i.e., Keating, 2008) and studies where participants were told that the purpose was vocabulary learning (i.e., Maftoon & Haratmeh, 2013). Additionally, we excluded studies where participants engaged in deliberate vocabulary learning conditions (e.g., word card learning, the keyword technique).
2. Studies that tested the prediction of the ILH and studies that coded IL for all learning conditions were included. Studies mentioning the ILH that did not clearly code each learning condition according to the ILH were excluded.
3. Studies that reported enough descriptive statistics to analyze posttest scores (i.e., the number of participants tested, mean, and SD for test scores) were included.
4. We excluded studies including a learning condition where multiple language activities were employed. The reason for this is that it is not clear how each component of the ILH contributed to learning gains when participants engage in multiple tasks involving different ILs.
5. Studies were excluded when their results were already reported in other publications that were included in our literature search.
6. Studies written in a language other than English were excluded.
7. Studies were excluded when activities were not described clearly enough to double-check the reported coding of the ILH. For instance, some studies reported that participants had to understand the target words in certain learning conditions but did not report how participants might learn the meanings of target words. We also excluded studies that failed to report how learning gains were measured and scored. This criterion also worked as a gatekeeper to ensure the quality of the included studies, especially because we included non-peer-reviewed studies as well as peer-reviewed studies.

The abstracts of the research reports identified through the literature search were carefully examined, and full texts were retrieved for 137 studies that examined vocabulary learning and mentioned the ILH. Through further examination, we found 40 studies meeting all of our criteria. Furthermore, we contacted the authors of 14 other studies that were only lacking in the descriptive statistics and gratefully received information from two authors (Hazrat, 2015; Tang & Treffers-Daller, 2016). Overall, 42 studies (N = 4628) that reported 398 mean posttest scores met all of our inclusion and exclusion criteria. These

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

included studies were 30 journal articles, 4 master's theses, 3 book chapters, 2 doctoral dissertations, 2 conference presentations, and 1 bulletin article (see Appendix S1 in Supporting Information online for basic information about the studies).[2]

DEPENDENT VARIABLE: EFFECT SIZE CALCULATION

To analyze the reported posttest scores on a standardized scale, we followed earlier meta-analyses on vocabulary research (Swanborn & de Glopper, 1999; Yanagisawa et al., 2020) and calculated the proportion of unknown target words learned (a.k.a. relative learning gain; Horst et al., 1998) as an effect size (ES).

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

Similarly, sampling variances of the posttest scores were calculated from reported SDs after converting them into the same scale using the escalc function of the metafor package (Viechtbauer, 2010) in the R statistical environment (R Core Team, 2017). Each calculated ES was weighted using the sampling variance of the posttests scores (see Appendix S2 in Supporting Information online for the detailed calculation formulas for ES and sampling variance).

PREDICTOR VARIABLES

We coded the studies for predictor variables: the ILH components, test format, test day (i.e., the number of days between learning and testing), frequency, mode, and number of target words (see Appendix S3 in Supporting Information online for the details on the coding scheme used).

Involvement Load Hypothesis components

The IL for each learning condition was coded strictly following Laufer and Hulstijn's (2001) description of the ILH. Learning conditions were coded for each ILH component (need, search, and evaluation) as either (a) absent, (b) moderate, or (c) strong. Using this predictor variable, we allow each component (and its levels) to contribute to learning gains to different degrees.

Additionally, different operationalizations of the ILH were adopted to code learning conditions. We coded learning conditions to distinguish two different types of strong evaluation (a) when each target word was used in a sentence (e.g., sentence writing) and (b) when a set of target words were used in a composition (written passages including multiple sentences, e.g., composition-, summary-, and letter-writing). To distinguish between the different levels of evaluation more clearly, we relabeled the levels: no evaluation, evaluation (i.e., comparison of words or meanings), sentence-level varied use (i.e., using a word in a sentence), and composition-level varied use (i.e., using a word in a composition).

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

Test format

Test format was coded as either (a) meaning recognition, (b) form recognition (meaning cue), (c) form recognition (form cue: select the appropriate spellings of target words; Martínez-Fernández, 2008), (d) meaning recall, (e) form recall, (f) vocabulary knowledge scale (VKS; e.g., Wesche & Paribakht, 1996), or (g) use of target words—fill-in-the-blanks (e.g., Jahangard, 2013) or writing (participants were asked to use a word in a sentence with grammatical and semantic accuracy; e.g., Bao, 2015). Three different groupings were then prepared: (a) each test format (i.e., each test format was grouped separately), (b) recall (meaning recall and form recall) versus recognition (meaning recognition and form recognition) versus other (VKS and use of target words), (c) receptive (receptive recognition and receptive recall) versus productive (form recognition and form recall) versus other (VKS and use of target words), and (d) receptive recall versus productive recall versus recognition versus other (VKS and use of target words).

Other predictor variables

The number of days between learning and testing was coded as test day. Frequency was coded for the number of times participants encountered or used each target word during a task. Mode was coded as either (a) written when participants engaged in a written activity (i.e., reading and writing) or (b) spoken when participants engaged in a spoken activity (i.e., listening and speaking; e.g., Hazrat, 2015; Jahangard, 2013). Lastly, the number of target words that participants were exposed to during a task was coded.

To ensure the reliability and consistency of the coding, four researchers were involved in the coding. First, one author of this meta-analysis, and another researcher who had carried out other meta-analyses and whose expertise included vocabulary research coded three studies separately using the developed coding scheme. There was no discrepancy across the two coders. All potential confusion was discussed, and the coding scheme was revised to make coding clearer and more objective. Next, one author carefully coded the 42 studies, and then 22 studies (52.4%) were randomly selected and double-coded separately by two other researchers in the field of applied linguistics who had also carried out meta-analyses. The intercoder reliabilities were calculated using Cohen's Kappa coefficient (κ) and the agreement rate was high and acceptable at $\kappa = .975$ and .987 for each double-coder. All discrepancies were resolved through discussion, and the first author again carefully double-checked the coding of all included studies to ensure consistency in coding.

DATA ANALYSIS

We used a three-level meta-regression model (Cheung, 2014; Lee et al., 2019) to analyze ESs that indicate the proportion of unknown words learned (Swanborn & de Glopper, 1999; Yanagisawa et al., 2020). Three-level meta-regression models can account for different sources of variance (i.e., within- and between-study variances and sampling variances), thus allowing sensible analyses of learning gains from different learning conditions compared within a study. Additionally, many studies reported more than one posttest score that were not independent (e.g., the same participants were tested repeatedly or with different test formats). To deal with this, the correlations across ESs from the same study were imputed to be 0.5 and applied to the analysis using the impute_covariance_matrix function of the clubSandwich package (Pustejovsky, 2017, 2018; see also, e.g., Teixeira-Santos et al., 2019 adopting a similar approach).[3] Furthermore, we adopted the cluster robust variance estimation (RVE) (Hedges et al., 2010) with small sample adjustments (Tipton, 2015; Tipton & Pustejovsky, 2015) when assessing the significance of the coefficients of predictor variables.

Three-level meta-regression models with maximum likelihood estimation were fitted with the rma.mv function of the metafor package (Viechtbauer, 2010) while specifying three different sources of variance: sampling variance of the effect sizes (level 1), variance between effect sizes from the same study (level 2, within-study variance), and variance across studies (level 3, between-study variance). ESs of immediate and delayed posttest scores were analyzed separately.

Analysis procedure

We used an information-theoretic approach to select the best predictive model from candidate models by referring to Akaike Information Criteria (AIC; Akaike, 1974; Burnham & Anderson, 2002). In this approach, statistical models including different predictor variables (or different combinations of predictor variables) were ranked by the model's AIC value. The model with the smallest AIC value has the greatest predictive power among all candidate models (Burnham & Anderson, 2002; see also Viechtbauer, 2020, for the application to meta-regression). Following Burnham and Anderson (2002), we used Akaike's Information Criterion corrected (AICc) for small sample sizes (Sugiura, 1978) as a reference.

To answer our research question, we first identified the best operationalization of the ILH and the best grouping of test formats, then determined the best combination of variables contributing to the prediction of incidental vocabulary learning. To identify the best operationalization of the ILH, three statistical models were fitted: (a) the original ILH model that only includes IL as a single numerical predictor variable (the sum of the scores of the three ILH components, a.k.a. task-induced IL index, Laufer & Hulstijn, 2001, p. 16; see also Hulstijn & Laufer, 2001, p. 544), (b) the ILH component model that includes categorical variables denoting each of the components (need, search, evaluation) separately for each level (absent, moderate, and strong), and (c) the modified ILH component model, which included the same predictor variables as the second model except for evaluation being four levels: no evaluation, evaluation, sentence-level varied use, and composition-level varied use. These three models are fitted with three-level meta-regression models and compared by their AICc values to determine the optimal operationalization of the ILH.

Similarly, we identified the best grouping of test formats using model selection with AICc. This was to best group the different test formats with similar sensitivities to learning so as to enhance the prediction of learning gains. While controlling the influence of IL by using the identified best ILH operationalization, we fitted four models based on the different groupings of test formats: (a) each test format; (b) receptive, productive, and other; (c) recall, recognition, and other; and (d) receptive recall, productive recall, recognition, and other.

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

Then, we conducted an automated model selection to determine the best predictive model that includes variables contributing to the prediction of learning gains. The models, including other potential predictor variables (i.e., frequency, number of target words, mode, plus test day for a model analyzing delayed posttests) as well as the identified best operationalization of the ILH and the grouping of test formats, were automatically analyzed with the glmulti package by comparing exhaustive combinations of all predictor variables while referring to AICc. Estimated coefficients were evaluated using an RVE with the clubSandwich package (Pustejovsky, 2018).

To evaluate whether the predictive power was enhanced from the original ILH, the explained variance was calculated at within- and between-study levels (Cheung, 2014) for the resulting model and the original ILH model that only included IL as a predictor variable. The explained variance at the within-study level indicates the proportion of explained variance in ESs across conditions within studies. This roughly corresponds to the variance explained by the framework while the effects of the characteristics of target words and participants are held constant. We also calculated the overall explained variance (the sum of the variance explained both at within- and between-study levels) so as to examine the explanatory power of each framework across studies. Because the present study did not include predictor variables that are specifically aiming to explain the variance across studies, the explained variance at the between-study level will not be interpreted. Because explained variance is nonnegative by definition, negative values were truncated and interpreted as zero (Cheung, 2014).

Lastly, sensitivity analyses were conducted to confirm the robustness of the obtained results (see Online Supplementary Appendix S4 in the Supporting Information online).

RESULTS

To identify the best combination of predictive variables for incidental vocabulary learning, we first compared different operationalizations of the ILH to determine which ILH operationalization best predicts learning gains. Three-level meta-regression models were fitted with three different operationalizations of the ILH: (a) an original ILH model that only included IL as a single numerical predictor variable (the sum of the scores of the three ILH components), (b) an ILH component model that included categorical variables denoting each ILH component (need, search, and evaluation) at each level (absent, moderate, and strong, with absent being the reference level), and (c) a modified ILH component model, where evaluation had four levels (absent, moderate evaluation, sentence-level varied use, and composition-level varied use) with other predictor variables being the same as the second model. Among the included studies, no study included learning conditions with strong need; thus, the need variable was either absent or moderate.

The results showed that the modified ILH component model was the best model as indicated by its smallest AICc value (–149.23 on the immediate posttest and –159.72 on the delayed posttest) followed by the ILH component (–147.27 and –166.01) and the original ILH (–139.81 and –158.79) in that order (see Table 1; note that these AICc values are all negative, thus the greater the number, the smaller the AICc value). The calculated Akaike weights also indicated strong support for the modified ILH component model, indicating that the probability that this model is the best predictive model among all

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

TABLE 1. Comparison of the different ILH operationalizations

	Immediate Posttest			Delayed Posttest		
Framework	AICc	ΔAICc	Akaike Weight	AICc	ΔAICc	Akaike Weight
Original ILH model	–139.81	–	0.01	–158.79	–	0.00
ILH component model	–147.27	–7.46	0.27	–166.01	–7.22	0.13
Modified ILH component model	–149.23	–9.42	0.72	–169.72	–10.93	0.86

Note: The smaller the AICc value the better the model; as the values are all negative, Modified ILH component model fitted the data best, followed by ILH component model, then Original ILH model. Akaike weight indicates the probability that each model is the best model among all candidate models.

candidate models was 72% on the immediate and 86% on delayed posttests (see e.g., Symonds & Moussalli, 2011 for Akaike weight).

Next, three-level meta-regression models comparing the four models of different test format groupings were fitted while specifying the identified best ILH operationalization—the modified ILH component model—as a covariate. The results showed that (a) when test formats were grouped as receptive recall, productive recall, recognition, and other, AICc value was the smallest (–201.03 on the immediate and –224.70 on the delayed posttests), which indicates that this is the grouping of test formats that best predicts learning gains. This grouping was followed by (b) each test grouping (–197.33, –220.72), (c) recall versus recognition versus other (–189.31, –209.82), and (d) receptive versus productive versus other (–163.97, –189.27), in that order (Table 2). This was also strongly supported by the calculated Akaike weights (86% on immediate and 88% on delayed posttests), which indicated that the probability that the model grouping test format as receptive recall, productive recall, recognition, and others was the best predictive model among all candidate models.

Lastly, to identify the best combination of variables to predict incidental vocabulary learning, we used the automated model selection specifying the identified optimal ILH operationalization and the optimal test format grouping, as well as the other candidate predictor variables (i.e., frequency, mode, test day, and the number of target words). Frequency and test day were included as numerical variables. Test day was only included for the delayed posttest. Mode had two levels (written and spoken) and written was set as

TABLE 2. Comparison of the different test format groupings while controlling ILs

	Immediate Posttest			Delayed Posttest		
Test grouping	AICc	ΔAICc	Akaike Weight	AICc	ΔAICc	Akaike Weight
Receptive vs. Productive vs. Other	–163.97	–	0.00	–189.27	–	0.00
Recall vs. Recognition vs. Other	–189.31	–25.34	0.00	–209.82	–20.55	0.00
Each test format	–197.33	–33.36	0.14	–220.72	–31.45	0.12
Receptive Recall vs. Productive Recall vs. Recognition vs. Other	–201.03	–37.06	0.86	–224.70	–35.43	0.88

Note: The smaller the AICc value the better the model; as the values are all negative, the last grouping (Receptive Recall vs. Productive Recall vs. Recognition vs. Other) fitted the data best. Akaike weight indicates the probability that each model is the best model among all candidate models.

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

the reference level. All predictor variables were analyzed with the glmulti package to compare models with exhaustive combinations of all predictor variables while referring to AICc. The resulting model with the smallest AICc will include the optimal combination of predictor variables that best predicts learning gains.

Table 3 and Table 4 show the optimal models selected for immediate and delayed posttests, respectively. The resulting model predicting L2 incidental vocabulary learning on immediate posttests included seven predictors: need, evaluation, sentence-level varied

TABLE 3. Parameter estimates and P-values for the predictor variables Included in the best model on the immediate posttest

		95% CI		
Predictor variables	Estimate	Lower	Upper	*p*
Intercept	0.074	–0.084	0.233	.334
Test: Productive recall	–0.127	–0.225	–0.028	.023
Test: Recognition	0.225	0.042	0.409	.035
Test: Other	–0.099	–0.158	–0.040	.009
Need	0.209	0.037	0.381	.024
Evaluation	0.083	0.039	0.126	.001
Varied Use (Sentence)	0.153	0.080	0.225	< .001
Varied Use (Composition)	0.233	0.131	0.335	< .001
Frequency	0.094	0.012	0.176	.033
Mode: Spoken	–0.098	–0.225	0.029	.091
Total explained variance	.168			
Between-study variance explained	.000			
Within-study variance explained	.590			

Note: 95% CIs and p-values were calculated based on the robust variance estimation. For reference level, test format was set as receptive recall, and mode was set as written.

TABLE 4. Parameter estimates and P-values for the predictor variables included in the best model on the delayed posttest

		95% CI		
Predictor variables	Estimate	Lower	Upper	*p*
Intercept	0.188	0.066	0.311	.006
Test: Productive recall	–0.123	–0.277	0.030	.090
Test: Recognition	0.192	–0.009	0.392	.054
Test: Other	–0.089	–0.132	–0.046	.004
Need	0.138	0.029	0.248	.021
Search	–0.051	–0.122	0.020	.140
Evaluation	0.091	0.046	0.137	.001
Varied Use (Sentence)	0.115	0.062	0.167	< .001
Varied Use (Composition)	0.210	0.157	0.262	< .001
Test day	–0.004	–0.007	–0.002	.014
Total explained variance	.339			
Between-study variance explained	.194			
Within-study variance explained	.562			

Note: 95% CIs and *p*-values were calculated based on the robust variance estimation. For reference level, test format was set as receptive recall.

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

use, composition-level varied use, test format, frequency, and mode. Search and the number of target words were not included in this model. The analyses of the variables related to the ILH components showed that need, evaluation, sentence-level varied use, and composition-level varied use, all positively contributed to learning. The estimated mean learning gain increased by 20.9% for the inclusion of need ($b = 0.209$, $p = .024$), 8.3% for evaluation ($b = 0.083$, $p = .001$), 15.3% for sentence-level varied use ($b = 0.153$, $p < .001$), and 23.3% for composition-level varied use ($b = 0.233$, $p < .001$). The analyses of test format revealed that with receptive recall being the reference level, when gains were measured with productive recall and “other” test formats, learning decreased by 12.7% and 9.9%, respectively. In contrast, when learning was measured with recognition tests, gains increased by 22.5%. The analyses also showed that learning gains increased by 9.4% as frequency increased by 1 and decreased by 9.8% when mode was spoken (as opposed to written).

All the included predictors’ influence were confirmed with 95% confidence intervals (CIs) and *p*-values calculated based on the RVE, except for mode ($p = .091$). Given that model selection and significance testing are two different analytic paradigms, the fact that mode was included in the model but did not reach the conventional statistical significance ($p < .05$) suggests that mode is a useful factor to predict learning gains although its influence may require further examination to confirm whether it is statistically significant or not (Burnham & Anderson, 2002; see also, Aho et al., 2014).

The resulting model on delayed posttests included seven predictors: need, search, evaluation, sentence-level varied use, composition-level varied use, test format, and test day. Frequency, mode, and the number of target words were not included in the model. The analysis of the variables related to the ILH components showed that need, evaluation, sentence-level varied use, and composition-level varied use all positively contributed to learning, except for search. The estimated mean learning gain increased by 13.8% for the inclusion of need ($b = 0.138$, $p = .021$), 9.1% for evaluation ($b = 0.091$, $p = .001$), 11.5% for sentence-level varied use ($b = 0.115$, $p < .001$), and 21.0% for composition-level varied use ($b = 0.210$, $p < .001$). The analyses of test format revealed that with receptive recall being the reference level, when gains were measured with productive recall and “other” test formats, learning decreased by 12.3% and 8.9%, respectively. Whereas, when learning was measured with recognition tests, learning increased by 19.2%.

Among the included predictors, test day and search were negatively related to learning gains. The results showed that learning gains decrease by 0.4% as the number of days between learning and testing increases by 1 ($b = –0.004$, $p = .014$). The results also showed that when search was present, the estimated mean learning gain decreased by 5.1% ($b = –0.51$, 95% CI [–0.122, 0.020]). Additionally, *p*-value calculated based on the RVE showed that search did not reach statistical significance ($p = .140$), suggesting that there is great variance in the negative influence of search and it might not necessarily hinder learning, but may be useful to include for prediction. To confirm that the negative influence of search is statistically significant or not, further investigation with larger sample sizes may be required.

The resulting models both on the immediate and delayed posttests also showed greater predictive power than the original ILH as indicated by the increased explained variance at the within-study level (i.e., the variance explained within the same study) and the total variance level (i.e., the sum of the variances at the within- and between-study levels

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

explained by the model) (Cheung, 2014). The original ILH model explained 15.3% of the total variance and 27.8% of the within-study variance on immediate posttests, and 5.8% of the total variance and 25.1% of the within-study variance on delayed posttests. The resulting model explained 16.8% of the total variance and 59.0% of the within-study variance on the immediate posttest, and 33.9% of the total variance and 56.2% of the within-study variance on delayed posttests.[4] The much greater explained variance provided by the resulting models indicates that they provide more accurate estimations of learning gains from incidental vocabulary learning activities than the original ILH.

DISCUSSION

The current study aimed to create a comprehensive framework of incidental vocabulary learning by meta-analyzing studies testing the effect of IL on incidental vocabulary learning. The optimal operationalization of the ILH (i.e., the modified ILH component model, where evaluation had four levels) and test format grouping (receptive recall vs. productive recall vs. recognition vs. other) were identified, then the automated model selection produced the resulting models that included a set of meaningful predictor variables.

The resulting models showed greater predictive ability, as indicated by the larger explained variance compared to the original ILH. The explained variance at the within-study level increased by 31.2% (from 27.8% to 59.0%) on immediate posttests and by 31.1% (from 25.1% to 56.2%) on delayed posttests. Given that the within-study variance reflects the learning gain differences among conditions within the same study, the same groups of participants, and using the same set of target words, this result suggests that the resulting statistical models provide a more accurate estimation of learning when other factors (i.e., test type, mode, test day, and frequency) were accounted for statistically. Furthermore, the total explained variance also increased by 1.5% (from 15.3% to 16.8%) on the immediate posttest and by 28.1% (from 5.8% to 33.9%) on the delayed posttest. This suggests that the resulting models predict learning gains better than the original ILH even when comparing the posttest scores across different learning situations where different groups of students are learning different sets of target words.

WHAT IS THE BEST COMBINATION OF PREDICTOR VARIABLES FOR INCIDENTAL VOCABULARY LEARNING?

In answer to the research question, the model selection approach identified the optimal combinations of predictors of incidental vocabulary learning within the meta-analyzed studies. The resulting models included the variables related to the ILH components, test format, and other empirically motivated variables (i.e., frequency, mode, and test day). The main conditions contributing to learning both on the immediate and delayed posttests were (a) need, (b) evaluation, (c) sentence-level varied use, and (d) composition-level varied use. As earlier studies argued (Kim, 2008; Laufer & Hulstijn, 2001), examining the contributions of the IL components on their own, rather than the combined IL components as a whole, significantly enhanced the prediction. Additionally, revising the evaluation component by distinguishing between different types of *strong evaluation* (i.e., sentence-level varied use and composition-level varied use) also led to a better model fit. One

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

plausible explanation for this is that learners benefit more from using a set of unknown words together in a text (e.g., a composition) compared to using each word in a separate sentence because using a set of words in a passage may elicit greater attention to how words can be used meaningfully. Another explanation may be that generating a text that coherently includes all target words induces pretask planning and hierarchal organization where learners must pay greater attention to the organization of the target words beforehand (Zou, 2017). Perhaps planning for the interaction with each word leads to greater learning.

The influence of test format was determined to be quite similar between the immediate and delayed posttests; recognition showed the highest gains, followed by receptive recall, other test formats (i.e., VKS, sentence-writing, gap-filling), and productive recall, in that order. With receptive recall being the reference, learning gains decreased when measured with productive recall (by 12.7% on immediate and by 12.3% on delayed posttests) and other test formats (by 9.9% and 8.9%) but increased with recognition (by 22.5% and 19.2%). Given that the type of test greatly influences learning gains (Webb, 2007, 2008), these results may be valuable when estimating overall learning gains. The present study also highlighted the value in comparing different groupings of measurements for finding optimal categorizations when creating a predictive model of learning.

Frequency and mode were also found to contribute to the prediction on the immediate posttest. The results showed that learning gains increased as frequency increased, corroborating earlier studies examining the effects of frequency and IL on vocabulary learning (Eckerth & Tavakoli, 2012; Folse, 2006). This highlights the importance of quantity as well as quality for word learning (Hulstijn, 2001; Schmitt, 2010; Webb & Nation, 2017). On immediate posttests, learning gains were estimated to increase by 9.4% as frequency increased by 1 and decrease by 9.8% when mode of input was spoken (as opposed to written). These findings provide useful pedagogical implications about how incidental vocabulary learning conditions might be improved. Learning may be increased by simply increasing the frequency of encounter or use of target words. Therefore, developing or selecting activities that involve multiple encounters or use of target items should be encouraged. The finding also advocates for the effectiveness of repeated-reading and -listening (Serrano & Huang, 2018; Webb & Chang, 2012) and narrow-reading, -listening, and -viewing in which repetition of target items is central to the activity (Chang, 2019; Rodgers & Webb, 2011). Similarly, having students engage in the same activities (or materials) including the same set of target words multiple times may also enhance vocabulary learning.

The finding for mode indicated that spoken activities (listening and speaking) tended to lead to lower learning gains than written activities (e.g., reading, writing, and gap-filling) when measured immediately after the learning session. This finding is supported by two earlier studies that indicated that reading leads to greater incidental vocabulary learning than listening (Brown et al., 2008; Vidal, 2011) but contrasted with another study that found no difference between the two modes (Feng & Webb, 2020). One reason why reading might contribute to learning to a greater extent than listening is that learners can pause, attend to words for as long as necessary, and even return to a word during a task using written input. In contrast, given the online nature of listening, spoken activities may provide a limited amount of time to attend to target words (Uchihara et al., 2019; Vidal, 2011). Another explanation could be that L2 learners tend to have a limited capacity for

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

processing L2 spoken input, limited phonological representations of L2 words (e.g., McArthur, 2003), and a smaller oral vocabulary than written vocabulary once their lexical proficiency develops to a certain level (Milton & Hopkins, 2006).

The predictive model for delayed posttests showed that test day and search, as well as need, evaluation, sentence-level varied use, composition-level varied use, and test formats were useful predictors. Learning gains were estimated to decrease by 0.4% as the number of days between learning and testing increases by 1. This small forgetting rate may be explained by the *testing effect* (e.g., Roediger & Karpicke, 2006). The majority of the studies included in this study administered both immediate and delayed posttests. Repeatedly testing the same words may have promoted retention of the words. It may be useful for future studies to examine the extent to which taking immediate posttests impacts delayed posttest scores to draw a more accurate estimation of the rate at which words are forgotten.

Interestingly, it was found that including search in an activity potentially hinders learning. When search was present, learning retention measured on delayed posttests were estimated to decrease by 5.1%. Yanagisawa and Webb's (2021) earlier meta-analysis of the ILH reported that the different operationalizations of search (i.e., the use of paper dictionaries, electronic dictionaries, or glossaries [paper glossaries and electronic glosses]) did not influence the effect of search and no positive influence of search was found. The negative influence of search may be explained by the learning conditions in the studies where search was present. When an activity included search, learners had to use other resources (e.g., dictionaries) to find information about target words. This extra cognitive task may deprive learners of time to learn the words because time is spent searching, for example, using a dictionary, rather than engaging with the target items. Some words may have even been ignored because the searching behavior, such as dictionary use, can be quite demanding for L2 students (Hulstijn et al., 1996). Alternatively, in activities without search, students were provided with information about target words (using glosses or glossaries). Therefore, they may have had more time and opportunities to attend to or process target words by having the forms and meanings of target items provided at their disposal.

In contrast to our hypothesis, the number of target words did not clearly contribute to the prediction of learning gains. To confirm this, we manually added the number of target words variable to the resulting models to determine its influence. The results showed that although there was a trend of a weak negative correlation between the number of target words and learning gains on the immediate posttest ($b = –0.003$, 95% CI [–0.011, 0.004]), the number of target words was not significantly related to learning gains on either the immediate ($p = .206$) or the delayed posttests ($b = 0.000$, 95% CI [–0.013, 0.014], $p =$ .929). One explanation may be that each study provided participants with sufficient time to complete the task given the difficulties related to the characteristics of target words, tasks, and participants as well as the number of target words. These findings may indicate that if learners can appropriately complete a task, a larger number of target words does not necessarily lead to lower learning gains.

Lastly, the resulting statistical models are slightly different between immediate posttests and delayed posttests. This points to the possibility that different factors influence immediate learning and retention in different manners. First, frequency and mode (i.e., the advantage of written over spoken activities) did not contribute to the prediction on delayed posttests while they did on immediate posttests. One plausible explanation is that the positive influences of increased frequency and the written mode fade over time because the retention of learned words declines as time passes. Especially for frequency, it may require multiple encounters or uses of a word over several days (as opposed to in a one-day learning session) for the word to be entrenched in memory (see Uchihara et al., 2019). Lastly, search negatively impacted learning only on delayed posttests. The presence of search—for example, a learner needs to take the time to consult with a dictionary as opposed to when a glossary is provided—potentially decreases the frequency to process or use target words during an activity. However, despite the distractive nature of search, the information about a word is indeed processed. Therefore, a learner may have been able to retrieve the word from memory when asked immediately after learning.

INCIDENTAL VOCABULARY LEARNING FORMULAS AND ILH PLUS

Based on the effect of predictors indicated by the resulting models, we created incidental vocabulary learning (IVL) formulas to estimate the relative effectiveness of different incidental learning tasks in a similar manner as the ILH. Two IVL formulas were created; one to estimate learning on immediate posttests and the other to estimate retention on delayed posttests.

$$\begin{aligned}
&\textit{The incidental vocabulary learning formula of activities for immediate learning}\\
&\qquad = [\textit{Need}\ (\textit{absent}\ :\ 0\ \textit{or present}\ :\ 1)\ \times\ 20.9]\\
&\qquad + [\textit{Evaluation}\ (0\ \textit{or}\ 1)\ \times\ 8.3]\\
&\qquad + [\textit{Sentence-level varied use}\ (0\ \textit{or}\ 1)\ \times\ 15.3]\\
&\qquad + [\textit{Composition-level varied use}\ (0\ \textit{or}\ 1)\ \times\ 23.3]\\
&\qquad + [\textit{Frequency}\ (\textit{number of time stone counter or use})\ \times\ 9.4]\\
&\qquad + [\textit{Mode}\ (\textit{writter}\ :\ 0\ \textit{or spoken}\ :1)\ \times\ -9.8]
\end{aligned}$$

$$\begin{aligned}
&\textit{The incidental vocabulary learning formula of activities for retention}\\
&\qquad = [\textit{Need}\ (\textit{absent}\ :\ 0\ \textit{or present}\ :\ 1)\ \times\ 13.8]\\
&\qquad + [\textit{Search}\ (0\ \textit{or}\ 1)\ \times\ -5.1]\\
&\qquad + [\textit{Evaluation}\ (0\ \textit{or}\ 1)\ \times\ 9.1]\\
&\qquad + [\textit{Sentence-level varied use}\ (0\ \textit{or}\ 1)\ \times\ 11.5]\\
&\qquad + [\textit{Composition-level varied use}\ (0\ \textit{or}\ 1)\ \times\ 21.0]
\end{aligned}$$

(See also Online Supplementary Appendix S5 in the Supporting Information online for explanations and examples of coding with the incidental vocabulary learning formulas.)

The formulas include seven components (need, search, evaluation, sentence-level varied use, composition-level varied use, frequency, and mode) to calculate the effectiveness index of an activity. The effectiveness index expresses the relative effectiveness of activities for incidental vocabulary learning; an activity with a higher effectiveness index is estimated to lead to larger vocabulary learning gains than another activity with a lower effectiveness index. For example, if the task is to write a composition using each of the target words three times, its effectiveness index is 72.4 (= 20.9 [need] + 0 [evaluation] + 0 [sentence-level varied use] + 23.3 [composition-level varied use] + 3 * 9.4 [frequency] + 0 [mode: spoken]) for immediate learning. This task is estimated to lead to greater vocabulary learning than a sentence-writing task using each word once, obtaining 45.6 (= 20.9 + 0 + 15.3 + 0 + 1 * 9.4 + 0) for its effectiveness index. In contrast, a similar sentence-writing task using each word six times is estimated to outperform both of the above tasks as it has an effectiveness index of 92.6 (= 20.9 + 0 + 15.3 + 0 + 6 * 9.4 + 0). Note that because none of the analyzed studies included learning conditions with strong need, need was included at two levels (absent or present).

Based on the proposed formulas, we propose an ILH Plus:

1. With other factors being equal (i.e., with the same test format at the same timing, the same set of target words, and dealing with the same population of participants), language activities with a higher effectiveness index calculated with the IVL formulas will lead to greater incidental word learning than activities with a lower effectiveness index.
2. Regardless of other factors that are not included in the IVL formulas, language activities with a higher effectiveness index will lead to greater incidental word learning than activities with a lower effectiveness index.

The first hypothesis may be useful for researchers to test ILH Plus as a falsifiable hypothesis to evaluate how accurately ILH Plus predicts the relative effectiveness of activities. This hypothesis corresponds to the original ILH's assumption (Laufer & Hulstijn, 2001). We do not necessarily claim that the values in the IVL formulas are the exact size of each factor's impact nor that the estimated task effectiveness will hold every single time. Instead, we claim that by testing the prediction of the formulas, researchers can approximate the magnitude of multiple variables, enhance our prediction of incidental vocabulary learning, and deepen our understanding of how multiple factors interact with each other to influence learning.

The second hypothesis is proposed as a null hypothesis. Researchers can examine whether the prediction of ILH Plus holds even when other variables as well as the factors of ILH Plus are manipulated. One example of a study testing Hypothesis 2 would explore learners' proficiency while manipulating the factors of the IVL formulas as in Kim (2008). By measuring each student's L2 proficiency level (e.g., which could be approximated by their vocabulary size) and manipulating the factors of the formulas, such a study can examine (a) whether the effect of proficiency overrides the estimation of the effectiveness index and (b) whether the predictive power of the effectiveness index varies based on students' proficiency. If the results would support Hypothesis 2 (e.g., by showing that different proficiency levels led to similar learning gains when controlling for the effectiveness index), this may suggest that proficiency may not influence learning to a meaningful extent. In contrast, if the results would reject Hypothesis 2, this may highlight the effect of proficiency, and the obtained data enables the analysis of the magnitude of proficiency effects and whether there is an interaction between proficiency and effectiveness index.

Similarly, through testing Hypothesis 2, individual studies can manipulate various factors, such as the characteristics of target words (e.g., cognates or noncognates,

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

pronounceability, imageability, and word length), learners (e.g., proficiency and vocabulary size), and learning conditions (e.g., reference language [L1 or L2 used for glossaries], different operationalizations of the components of ILH Plus, and underinvestigated ILH components [e.g., strong need]). Because various factors are controlled within the framework of ILH Plus, the effect of the examined factor can be synthesized in future meta-analyses.

Because we aimed to provide a formula to calculate the effectiveness index in the same manner as the original ILH, the IVL formulas only included the factors that are directly related to the learning conditions. Thus, the resulting statistical models' intercept, test format, and test day were not included because these factors are more closely related to how learning gains were measured and not related to learning conditions per se. Although these factors may be useful for calculating the estimated mean learning gains, they may not be suited for a hypothesis testing framework.

To illustrate how the proposed formulas can be used to estimate the relative effectiveness of tasks, activities in Laufer and Hulstijn (2001) and Kim (2008) were coded following the formula for immediate learning (see Table 5). The three activities in Laufer and Hulstijn (2001) were (a) reading with glosses, (b) fill-in-the-blanks, and (c) composition writing. All activities included need and frequency as 1. Effectiveness indices were calculated as 30.3 for reading with glosses, 38.6 for fill-in-the-blanks, and 53.6 for composition writing. When comparing the observed mean test scores in Laufer and Hulstijn (2001), ILH Plus correctly predicted that incidental learning gains were largest for composition writing, followed by fill-in-the-blanks, and lastly reading with glosses. Similarly, the four activities in Kim (2008) were also coded using the IVL formula. The effectiveness indices were calculated to be 30.3 for reading with glosses, 38.6 for fill-in-the-blanks, 53.6 for composition writing, and 45.6 for sentence writing. Among the 24 comparisons of the activities (6 comparisons across 4 activities × 2 test timing × 2 participant groups), the IVL formula correctly predicted 22 comparisons (91.7%) of the relative effectiveness between the activities.

One thing to note is that when the effectiveness indices between activities are similar to each other, the activities are more likely to lead to similar learning gains. For example, the effectiveness indices for reading with glosses and fill-in-the-blanks are 30.3 and 38.6, thus their difference in learning gains from these activities might be more difficult to detect compared to when comparing learning gains between activities that have greater differences in effectiveness indices such as composition writing (53.6) and reading with glosses (30.3). Because it is normal for mean scores to fluctuate, there will likely be times when the estimated rank order of effectiveness is not observed due to limited statistical power especially when the effectiveness index values are close across activities.

LIMITATIONS AND FUTURE DIRECTIONS

First, ILH Plus and the IVL formulas should be viewed as a simple predictive model. The proposed formulas are representative of the studies that were analyzed. However, these studies represent a limited set of possible tasks and learning contexts. For example, in earlier studies the effect of some predictive variables (i.e., frequency, mode, and search) were not extensively examined with different learning conditions that involve varying degrees of need, evaluation, and varied use (sentence- and composition-levels). Thus, it would be useful for

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

TABLE 5. Coding examples of the incidental vocabulary learning formula (immediate learning measured with immediate posttests)

	Hulstijn and Laufer (2001)			Kim (2008)			
	Reading with glosses	Fill-in-the-blanks in a text	Composition-writing	Reading with glosses	Fill-in-the-blanks in a text	Composition-writing	Sentence-writing
Need: × 20.9	1	1	1	1	1	1	1
Evaluation: × 8.3	0	1	0	0	1	0	0
Varied use (sentence): × 15.3	0	0	0	0	0	0	1
Varied use (composition): × 23.3	0	0	1	0	0	1	0
Frequency: × 9.4	1	1	1	1	1	1	1
Mode (Spoken): × −9.8	0	0	0	0	0	0	0
EI	30.3	38.6	53.6	30.3	38.6	53.6	45.6
Order of effectiveness	3	2	1	4	3	1	2

Note: EI = Effectiveness Index.

future studies to investigate the predictive accuracy of ILH Plus with learning conditions employing a greater variety of combinations of factors. Furthermore, the present study examined a limited numbers of predictor variables (e.g., ILH components, frequency, mode, test format, and test day). Although there are many other factors that potentially contribute to predicting learning gains (e.g., students' L2 proficiency, Kim, 2008; the characteristics of target words, Ellis & Beaton, 1993; gloss language, Laufer & Shmueli, 1997), these factors were not included in the analysis. This is because of the information-theoretic model selection approach adopted in the current study, which requires all predictor variables to be consistently reported. We encourage researchers in future studies to provide details on other factors such as proficiency information and gloss language (Uchihara et al., 2019; Yanagisawa et al., 2020) to allow further development of predictive models of vocabulary learning. To fully take advantage of the results of (quasi-) empirical studies, it would also be useful for future studies to make their materials (e.g., target words, glossaries, and reading texts) and datasets publicly available if possible. Having access to open materials and datasets enables more accurate coding and examination of a greater number of predictor variables.

Second, effects of interactions between factors were not included in ILH Plus or its formulas. This is mainly due to the limited combinations of factors investigated in the included studies. However, it might be reasonable to assume that the effect of a certain factor changes based on other factors. For example, the effects of varied use (both sentence- and composition-level) could be more pronounced when learning is measured with productive tests (e.g., form recall tests) compared to receptive tests (e.g., meaning recall). Similarly, the effect of some factors might increase or decrease based on other factors. For instance, the effect of frequency might be more pronounced when composition-level varied use was present compared to when evaluation was present. While search was found to negatively influence learning retention, there might be situations in which the positive effect of search can be observed. For example, search could influence learning positively when frequency increases because multiple encounters may provide students with retrieval opportunities (Nation & Webb, 2011). It would be useful for future studies to research different combinations of factors to examine how these variables interact with each other to influence incidental vocabulary learning. We hope that ILH Plus serves as a guideline for future studies to strictly control multiple variables so that each effect and their interaction effects can be easily examined.

Third, the current study identified some underresearched factors related to the ILH components. None of the meta-analyzed studies included learning conditions with strong need (internal motivation). Additionally, search was operationalized only as dictionary use, glossaries, electronic dictionaries, and hyperlinked glosses, with no studies examining situations in which students guess the meanings of words from context or ask teachers or peers. Furthermore, the number of studies investigating spoken activities was relatively small as the majority of studies used written activities. Therefore, the currently proposed ILH Plus is limited in these regards. Future studies should examine these underinvestigated conditions to further validate the original ILH and potentially improve ILH Plus.

Lastly, as one of the aims of the ILH was to provide a tool that helps language teachers and material developers enhance efficacy in language education (Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001), it may be important to consider the ecological validity of a task. As one limitation of this meta-analysis, we could not determine how suitable tasks and materials were for students in each study. The ILH studies tended to focus only on vocabulary learning

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

gains, whereas students' performance during and after an activity was rarely assessed. Rott (2012) tested the ILH while measuring participants' comprehension of a text used during activities and discussed the efficacy of the activities not only from vocabulary learning perspective but also from a communicative activity perspective. Similarly, assessing students' writing products may reveal how meaningful the created passages are. Measuring the performance of an activity as well as vocabulary learning may deepen our discussion of the efficacy of activities and their applicability to educational contexts.

CONCLUSION

We aimed to enhance the prediction of incidental vocabulary learning by meta-analyzing studies examining the ILH. The resulting statistical models show that the predictive power of the ILH was improved by (a) examining the influence of each level of individual ILH components and by (b) adopting the optimal operationalization of the ILH components and test format grouping. Including other empirically motivated variables also increased the prediction of the resulting model. Although ILH Plus may not provide 100% accurate predictions, it should serve as a more reliable tool than the original ILH, and one that language teachers, curriculum writers, and material designers can apply to their practice, as Box's oft-cited quotation notes "all models are wrong, but some are useful" (Box & Draper, 1987, p. 424).

Echoing Laufer and Hulstijn (2001), we would like to call for studies to examine the extent to which ILH Plus accurately predicts incidental vocabulary learning gains from language activities. Empirical studies can compare different learning conditions to determine whether ILH Plus accurately predicts incidental vocabulary learning. Specifically, studies might examine (a) whether the estimated order of the effectiveness of activities is as predicted and (b) whether the size of the contribution of each factor is as predicted. This can be realized not only with empirical studies strictly controlling other factors but also with classroom research examining how reliable ILH Plus is when applied to actual learning contexts. Studies might also investigate other factors that are not included in ILH Plus. These factors might include learner characteristics (e.g., proficiency, Kim, 2008; working memory, Yang et al., 2017), task covariates (e.g., time on task, Keating, 2008), lexical items (e.g., collocations, Snoder, 2017), reference language (e.g., gloss language, Laufer & Shmueli, 1997; Yanagisawa et al., 2020), and the similarity between learning and testing (transfer-appropriate-processing, Lightbown, 2008). Examining these factors while controlling the components of ILH Plus helps to (a) assess the relative effects of various factors, (b) investigate how multiple factors interact with each other to influence incidental vocabulary learning, and (c) build more complex models explaining the effectiveness of language activities. Lastly, after accumulating studies that test ILH Plus, meta-analyses can statistically summarize the findings of these studies to improve ILH Plus.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:
Appendix S1. Basic information about the included studies.
Appendix S2. Detailed calculation formulas for ES and sampling variance.

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

Appendix S3. Details of the coding scheme used.
Appendix S4. Sensitivity analyses.
Appendix S5. Explanations and examples of coding with incidental vocabulary learning formulas
Appendix S6. References of included studies.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S0272263121000577.

NOTES

[1] In this article, activity and task were used interchangeably to refer to instructional language activities that promote students' L2 vocabulary learning.

[2] The included studies in the analyses were the same as Yanagisawa and Webb (2021), and the dataset was greatly overlapped with that of Yanagisawa and Webb (2021).

[3] It is reasonable to assume that the ESs from the same study were correlated to some extent. However, none of the studies reported correlations between the scores of different test formats. The robustness of the correlation imputation was confirmed with additional analyses with varying correlations ($r = 0, .3, .5,$ and $.7$), which showed that the estimated effects of the predictor variables did not differ much and indicated the robustness of the results. If future studies make their entire dataset publicly available, future meta-analyses can calculate more accurate correlations between effect sizes to further enhance the accuracy of analysis.

[4] The slight difference between the variance explained by the original ILH reported in the current study and that reported in Yanagisawa and Webb (2021) is solely due to the fact that the current study specified the imputed correlations across the ESs obtained from the same study, whereas such imputation was not used in Yanagisawa and Webb (2021).

REFERENCES

The full reference list of the studies included in the meta-analysis is available in Appendix S6 in the Supporting Information online.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*, 631–636. https://doi.org/10.1890/13-1452.1

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. https://doi.org/10.2989/16073614.2016.1201774

Bao, G. (2015). Task type effects on English as a foreign language learners' acquisition of receptive and productive vocabulary knowledge. *System*, *53*, 84–95. https://doi.org/10.1016/j.system.2015.07.006

Barclay, S., & Schmitt, N. (2019). Current perspectives on vocabulary teaching and learning. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *Second handbook of information technology in primary and secondary education* (pp. 799–819). Springer. https://doi.org/10.1007/978-3-319-58542-0_42-1.

Beal, V. (2007). *The weight of involvement load in college level reading and vocabulary tasks* [Unpublished master's thesis]. Concordia University.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer-Verlag.

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press

Chang, A. C.-S. (2019). Effects of narrow reading and listening on L2 vocabulary learning: Multiple dimensions. *Studies in Second Language Acquisition*, *41*, 769–794. https://doi.org/10.1017/S0272263119000032

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

Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives* (First edition). Routledge

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

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research*, *16*, 227–252. https://doi.org/10.1177/1362168811431377

Ellis, N. C., & Beaton, A. (1993). *Psycholinguistic determinants of foreign language vocabulary learning. Language Learning*, *43*, 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

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. https://doi.org/10.1017/S0272263119000494

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*, 273–293. https://doi.org/10.2307/40264523

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*, 483–517. https://doi.org/10.1017/S0272263113000119

Hazrat, M. (2015). The effects of task type and task involvement load on vocabulary learning. *Waikato Journal of Education*, *20*, 79–92. https://doi.org/10.15663/wje.v20i2.189

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. https://doi.org/10.1002/jrsm.5

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.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal*, *96*, 544–557. https://doi.org/10.1111/j.1540-4781.2012.01394.x

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 University Press.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*, 327–339. https://doi.org/10.2307/329439

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*, 539–558. https://doi.org/10.1111/0023-8333.00164

In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly*, *44*, 169–184. https://doi.org/10.5054/tq.2010.215253

In'nami, Y., Koizumi, R., & Tomita, Y. (2020). Meta-analysis in applied linguistics. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 240–252). Routledge. https://doi.org/10.4324/9780367824471-21

Jahangard, A. (2013). Task-induced involvement in L2 vocabulary learning: A case for listening comprehension. *Journal of English Language Teaching and Learning*, *12*, 43–62.

Karalik, T., & Merç, A. (2016). The effect of task-induced involvement load on incidental vocabulary acquisition. *Mustafa Kemal University Journal of Graduate School of Social Sciences*, *13*, 77–92.

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. https://doi.org/10.1177/1362168808089922

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, *58*, 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

Kolaiti, P., & Raikou, P. (2017). Does deeper involvement in lexical input processing during reading tasks lead to enhanced incidental vocabulary gain? *Studies in English Language Teaching*, *5*, 406–428. https://doi.org/10.22158/selt.v5n3p406

https://doi.org/10.1017/S0272263121000577 Published online by 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., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*, 1–26. https://doi.org/10.1093/applin/22.1.1

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*, 89–108. https://doi.org/10.1177/003368829702800106

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, *40*, 721–753. https://doi.org/10.1093/applin/amy012

Lee, Y.-T., & Hirsh, D. (2012). Quality and quantity of exposure in L2 vocabulary learning. In D. Hirsh (Ed.), *Current perspectives in second language vocabulary research* (pp. 79–116). Peter Lang AG. https://doi.org/10.3726/978-3-0351-0379-3

Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han & E. S. Park (Eds.), *Understanding second language process* (pp. 27–44). Multilingual Matters.

Maftoon, P., & Haratmeh, M. S. (2013). Effects of input and output-oriented tasks with different involvement loads on the receptive vocabulary knowledge of Iranian EFL learners. *IJRELT*, *1*, 24–38.

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

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

McArthur, T. (2003). English as an Asian language. *English Today*, *19*, 19–22. https://doi.org/10.1017/S0266078403002049

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, *63*, 127–147. https://doi.org/10.3138/cmlr.63.1.127

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

Newton, J. (2020). Approaches to learning vocabulary inside the classroom. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 255–270). Routledge. https://doi.org/10.4324/9780429291586-17

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. https://doi.org/10.1017/S0267190510000115

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading. *Studies in Second Language Acquisition*, *38*, 97–130. https://doi.org/10.1017/S0272263115000224

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

Pustejovsky, J. E. (2017, August 10). *Imputing covariance matrices for meta-analysis of correlated effects.* https://www.jepusto.com/imputing-covariance-matrices-for-multi-variate-meta-analysis/

Pustejovsky, J. (2018). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections* (0.3.1) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

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

Rodgers, M. P., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*, 689–717.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

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

Sadoski, M. (2005). A dual coding view of vocabulary learning. *Reading & Writing Quarterly*, *21*(3), 221–238. https://doi.org/10.1080/10573560590949359

Sadoski, M., & Paivio, A. (2013). *Imagery and text: A dual coding theory of reading and writing* (2nd ed.). Routledge.

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

Serrano, R., & Huang, H.-Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, *52*, 971–994. https://doi.org/10.1002/tesq.445

Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of involvement load, spacing, and intentionality. *TESL Canada Journal*, *34*, 140–164. https://doi.org/10.18806/tesl.v34i3.1277

Sugiura, N. (1978). Further analysts of the data by Akaike's Information Criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics: Theory and Methods*, *7*, 13–26. https://doi.org/10.1080/03610927808827599

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

Symonds, M. R. E., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*, 13–21. https://doi.org/10.1007/s00265-010-1037-6

Tang, C., & Treffers-Daller, J. (2016). Assessing incidental vocabulary learning by Chinese EFL learners: A test of the involvement load hypothesis. In *Assessing Chinese learners of English* (pp. 121–149). Springer.

Teixeira-Santos, A. C., Moreira, C. S., Magalhães, R., Magalhães, C., Pereira, D. R., Leite, J., Carvalho, S., & Sampaio, A. (2019). Reviewing working memory training gains in healthy older adults: A meta-analytic review of transfer for cognitive outcomes. *Neuroscience & Biobehavioral Reviews*, *103*, 163–177. https://doi.org/10.1016/j.neubiorev.2019.05.009

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

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

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

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*, 219–258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

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

Viechtbauer, W. (2020, June 17). *Model selection using the glmulti and MuMIn packages [The metafor Package].* The metafor Package: A Meta-Analysis Package for R. http://www.metafor-project.org/doku.php/tips:model_selection_with_glmulti_and_mumin

Wang, C., Xu, K., & Zuo, Y. (2014). The effect of evaluation factor on the incidental vocabulary acquisition through reading. *International Journal of English Linguistics*, *4*, 59–66. https://doi.org/10.5539/ijel.v4n3p59

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

Webb, S., & Chang, A. (2012). Vocabulary learning through assisted and unassisted repeated reading. *Canadian Modern Language Review*, *68*, 267–290. https://doi.org/10.3138/cmlr.1204.1

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

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*, 13–40. https://doi.org/10.3138/cmlr.53.1.13

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental l2 vocabulary learning? A meta-analysis. *Language Learning*, *71*, 121–145. https://doi.org/10.1111/lang.12444

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

Yang, Y., Shintani, N., Li, S., & Zhang, Y. (2017). The effectiveness of post-reading word-focused activities and their associations with working memory. *System*, *70*, 38–49. https://doi.org/10.1016/j.system.2017.09.012

Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research*, *21*, 54–75. https://doi.org/10.1177/1362168816652418

https://doi.org/10.1017/S0272263121000577 Published online by Cambridge University Press