11-3-2020 11:00 AM

How can we Predict Incidental L2 Vocabulary Learning? A Meta-Analytic Examination of the Involvement Load Hypothesis

Akifumi Yanagisawa, *The University of Western Ontario*

Supervisor: Webb, Stuart A., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Education

Abstract

This dissertation investigated Laufer and Hulstijn's (2001) Involvement Load Hypothesis (ILH). The ILH claims that the retention of unknown words is conditional on one motivational factor (*need*) and two cognitive factors (*search* and *evaluation*) and predicts the relative effectiveness of activities on incidental vocabulary learning. While research tends to provide general support for the ILH, several studies revealed that the ILH prediction is not always accurate. Aiming to provide a summative evaluation of the ILH and enhance its predictive ability, the present thesis conducted a series of three meta-analytic studies to examine research that tested the ILH.

Chapter 1 outlines the thesis and provides background literature and the rationales for the three studies. Chapter 2 (Study 1) meta-analyzed studies testing the prediction of the ILH to investigate (a) the overall predictive ability of the ILH, (b) the relative effects of different components of the ILH, and (e) the influence of potential factors moderating learning. The results showed that the ILH significantly predicted learning gains. However, each ILH component contributed to learning differently and other factors were found to influence learning, suggesting potential for the ILH to be enhanced.

Chapter 3 (Study 2) aimed to update the ILH to enhance its accuracy in predicting learning. The results of the ILH studies were examined with the information-theoretic approach to determine the optimal statistical model that best predicts learning gains. The results showed that the prediction of the ILH improved by adopting the best operationalization of ILH components and optimal test format grouping and including other empirically motivated variables.

Chapter 4 (Study 3) systematically analyzed incidental vocabulary learning conditions that have been examined in studies of the ILH and calculated the estimated learning gains occurring across different activity types. The results revealed that the estimated mean learning gains were highest for composition-level varied use activities (e.g., composition-writing), followed by sentence-level varied use (e.g., sentence-writing), evaluation (e.g., fill-in-the-blanks), meaning-focused input (MFI; reading and listening) with need for comprehension of target words, and MFI in that order.

Lastly, Chapter 5 provides a final discussion of the thesis, followed by the limitations and potential future directions.

Keywords

Involvement Load Hypothesis, Incidental Vocabulary Learning, Depth of Processing, Meta-analysis, Predictive Modeling

Summary for Lay Audience

Anyone learning a new language must acquire an extensive vocabulary to develop a proficient command of that language. Therefore, second language (L2) teachers must choose language activities that effectively increase students' vocabulary. Laufer and Hulstijn's (2001) Involvement Load Hypothesis (ILH) is a framework that serves as a guide with which teachers can select activities that are effective for vocabulary learning. The ILH claims that L2 vocabulary learning is conditional on three factors: *need* (the necessity to understand or use a word), *search* (to look for information about a word), and *evaluation* (the comparison of the information about word meanings or forms). The level of presence of these components within an activity is called *Involvement Load* (IL), and the ILH predicts that language activities with higher ILs lead to greater vocabulary learning than activities with lower ILs.

Many studies have tested how accurately the ILH predicts the relative effectiveness of activities on vocabulary learning. While some studies report that their results supported the predictions of the ILH (e.g., Hulstijn & Laufer, 2001; Kim, 2008), other studies report that the ILH predictions were not always accurate (e.g., Folse, 2006; Keating, 2008). Aiming to provide a summative evaluation on the ILH and enhance its predictive ability, the present thesis examined the results of studies that tested the ILH by carrying out three studies.

The first study statistically summarized studies testing the prediction of the ILH to investigate how accurately the ILH predicts incidental vocabulary learning and how different

variables influence learning. The results showed that the ILH adequately predicted learning gains. However, the results also revealed some potential for the ILH to be enhanced.

The second study aimed to enhance the accuracy of the prediction of the ILH. The results showed that the prediction of the ILH improved by revising the operationalization of the ILH and including other variables.

The third study systematically overviewed incidental vocabulary learning conditions that have been examined in studies of the ILH. The learning conditions were grouped into five activity types, and we calculated the estimated learning gains for each activity type.

Co-Authorship Statement

Three studies from this dissertation have been submitted to research journals. Study 1 was submitted to *Language Learning* (22/Jan/2020), Study 2 to *Studies in Second Language Acquisition* (11/Aug/2020), and Study 3 to *the Modern Language Journal* (24/Aug/2020). All studies were co-authored with my supervisor, Dr. Stuart Webb.

I carried out data collection and data analysis, created the figures and tables, and wrote the original draft for each study. Dr. Stuart Webb provided supervision throughout the project. He also reviewed the papers and provided feedback prior to journal submission. Both authors contributed to the conceptualization, research design, and interpretation of results.

Acknowledgments

I would like to express my deepest appreciation to my supervisor, Stuart Webb, for his valuable guidance and his constant encouragement. He has been a great mentor for me and has helped me develop as a researcher. My appreciation also goes to Batia Laufer, Jan Hulstijn, Frank Boers, Judit Kormos, Emma Marsden, and the anonymous reviewers of Language Learning for their useful suggestions on this research project, and Akira Murakami, James E. Pustejovsky, Wolfgang Viechtbauer, Elizabeth Tipton, Mike Cheung, and Shusaku Kida for their input and knowledge about statistical analysis. I would also like to thank Takumi Uchihara, Tomlin Gagen, and Su Kyung Kim for their support in the process of data collection. Special thanks are due to Emi Iwaizumi, Zhouhan Jin, Yanxue Feng, Niousha Pavia, Juliane Martini, and other staff and colleagues at the University of Western Ontario.

Additionally, I wish to thank Hideki Sakai for his keen insight, warm encouragement, and showing me the attractiveness of research during my Master's program. I am also grateful to Kosuke Tanaka, Haruko Masuo, and Michiyo Nishimaki for their support regarding the Japan Student Service Organization (JASSO) scholarship as my PhD program was financially supported by JASSO.

My gratitude also goes to the following researchers who provided information about their studies for the current meta-analysis project: Jeanine Treffers-Daller, Chizuru Mori, Karim Jahangiri, Mandana Hazrat, Ali Jahangard, Mouhammad Reza Sarbazi, Mark Feng Teng, and Sasan Baleghizadeh.

I would like to express my sincerest gratitude to my parents, Naomi and Toshimi, my brothers, Naoe and Ryo, and my in-laws, Rich, Miki, and Joshua for their kindness, care, and continuous support.

Finally, my most heartfelt appreciation goes to my lovely wife, Kaitlyn, for her kindness, love, and consistent support and encouragement. She has been at my side from the beginning and enriches my life in many ways.

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 4

List of Abbreviations

AIC	Akaike Information Criteria
CI	Confidence Interval
ERIC	Educational Resources Information Centre
ES	Effect Size
L2	Second Language
LLBA	Linguistics and Language Behavior Abstract
ICC	Intra-class Correlation
IL	Involvement Load
ILH	Involvement Load Hypothesis
MFI	Meaning-focused Input
RVE	Robust Variance Estimation
TFA	Technique Feature Analysis
TOPRA	Type of Processing – Resource Allocation
VSK	Vocabulary Knowledge Scale

List of Appendices

Chapter 1

1 General Introduction

To develop a proficient command of a language requires an extensive vocabulary (e.g., Schmitt, 2008). It is therefore important for second language (L2) teachers to select language activities that effectively increase students' vocabulary. Laufer and Hulstijn's (2001) Involvement Load Hypothesis (ILH) is a framework that serves as a guide for language teachers to select activities that promote vocabulary learning. The ILH suggests that when learners pay more attention to unknown words and process words in an elaborated manner, these words are more likely to be recalled later. The ILH claims that the retention of new L2 words is contingent upon an activity's *Involvement Load* (IL), i.e., the extent to which learning conditions include three components: one motivational component (*need*, the necessity to understand or use a word) and two cognitive components (*search*, to look for information about a word, and *evaluation*, the comparison of information about word meanings or forms). The ILH predicts that language activities with higher ILs lead to greater vocabulary learning than activities with lower ILs.

Many studies have tested how accurately the ILH predicts the relative effectiveness of activities on incidental vocabulary learning. While some studies report that their results supported the predictions of the ILH (e.g., Eckerth & Tavakoli, 2012; Hulstijn & Laufer, 2001; Kim, 2008; Kolaiti & Raikou, 2017; Laufer, 2003), other studies report that the ILH predictions were not always accurate (e.g., Bao, 2015; Folse, 2006; Keating, 2008; Rott, 2012; Zou, 2017). Because of the inconsistency in studies, it is difficult to evaluate the extent to which the ILH predicts incidental vocabulary learning just by considering the findings of individual studies. Therefore, in order to examine the overall validity of the ILH, there is a need to systematically summarize earlier studies testing the ILH.

The present thesis meta-analyzed the results of (quasi-) empirical studies investigating the ILH to obtain a summative evaluation as to the extent to which the ILH

accurately predicts incidental vocabulary learning gains. Additionally, earlier findings point to potential directions to revise the ILH to improve its predictive ability. These suggestions include (a) adding other variables into consideration (frequency in Folse, 2006; time on task in Keating, 2008), (b) examining the influence of the individual components of the ILH (need, search, evaluation), and (c) revising the ILH components (e.g., distinguishing different types of evaluation, Zou, 2017). Based on these suggestions, the present thesis examines whether it is possible to enhance the predictive ability of the ILH. Lastly, the studies examining the ILH investigated vocabulary learning gains from a variety of learning conditions. Through meta-analyzing the results of these studies, the present thesis obtains the estimated learning gains for different activity types. Such estimated learning gains may produce transparent pedagogical implications, with which language teachers, learners, and curriculum writers can easily apply the accumulated research findings to practice without a deep understanding of vocabulary research.

1.1 Incidental Vocabulary Learning

Research has consistently demonstrated that second language (L2) students can learn vocabulary incidentally (Webb, 2020). Studies have revealed that vocabulary learning occurs through reading (Horst, Cobb, & Meara, 1998; Waring & Takaki, 2003), listening (e.g., Pavia, Webb, & Faez, 2019; van Zeeland & Schmitt, 2013), and viewing (e.g., Rodgers & Webb, 2019). Moreover, in addition to these meaning-focused input (MFI) activities, studies have also shown that students learn vocabulary as a by-product of completing a variety of language activities such as gap-filling (e.g., Kim, 2008, Folse, 2006), composition writing (Laufer, 2003), and sentence writing (e.g., Kim, 2008; Folse, 2006). Given that there are many activities, it is important for language teachers to select the most effective activities for vocabulary learning (Nation, 2007). To predict the relative efficacy of incidental learning activities, Laufer and Hulstijn (2001) proposed the *Involvement Load Hypothesis* (ILH).

1.2 Involvement Load Hypothesis

The ILH was developed to improve on earlier theories that aim to explain how quality of attention and cognitive processing of information influence memory retention. The concept of *depth of processing* (or *levels of processing*, Craik & Lockhart, 1972; Craik & Tulving, 1975) was perhaps the best known theory of vocabulary learning. Depth of processing argues that memory retention is conditional on how deeply an item is processed. The more deeply one processes a to-be-learned item, the more likely he or she can recall the item later. For example, Craik and Lockhart (1972) suggest that focusing on learning the semantic aspects of a word leads to deeper processing than focusing on learning the formal aspects of that word (e.g., spelling). The results from Craik and Tulving's (1975) study supported this hypothesis by showing that participants recalled more words in a condition that focused on learning the meanings of the target words when compared with a condition that involved learning the forms of the target words. However, the concept of 'depth of processing' faced two main criticisms: (1) it is ambiguous as to what exactly constitutes "depth" of processing, and because of this, (2) it is difficult to tell whether one task provides deeper processing than another (Baddeley, 1978; Craik & Tulving, 1975; Eysenck, 1978; Laufer & Hulstijn, 2001; Nelson, 1977). Laufer and Hulstijn (2001) developed ILH with an aim to create a framework to better predict L2 incidental vocabulary learning. They argued that vocabulary retention is contingent upon task-induced involvement load, which is determined by one motivational factor (*need*) and two cognitive factors (*search* and *evaluation*). The ILH provided clear criteria to evaluate language activities. By looking at the presence or absence of these three features (*need*, *search*, and *evaluation*) in a task, the likely effectiveness of the task on vocabulary learning could be predicted.

Need is a motivational factor and refers to whether the unknown word is needed to complete the task. It has three levels. When the unknown word is not required to complete the task, there is no need. Need is moderate when an external agent (e.g., a task or a teacher) asks learners to understand or use the word. Need is strong when the necessity of the word is generated by learners themselves. For example, need is absent when learners read a text and they encounter an unknown word, but that word is not

necessary to comprehend the text. Need is moderate when learners read a text and answer comprehension questions that require learners to understand the meaning of the unknown word because need is imposed by an external agent. Need is strong when learners read a book for pleasure and use a dictionary to look up unknown words to understand the story because the need to learn the words is generated by the learners themselves.

Search is a cognitive factor that refers to the attempt to find the meaning of an unknown word or the word itself to express a certain concept. When meaning and the word form are provided in the activity, there is no search. Search exists when the learners need to search for the meanings of the unknown words by consulting other authorities (e.g., a dictionary or a teacher). The search component exists in only one degree: absent or present. For example, search is absent when learners read a text with the meanings of unknown words provided in marginal glosses because learners do not need to search for the meanings of the words. Search is present when learners read a text while looking up the meanings of unknown words using a dictionary because learners need to search for the meanings of words by using other authorities. Search is also present when learners write a composition using unknown target words if only the word forms of the target words were listed and learners need to use a dictionary to look up the meaning of each word. Originally, the search factor was either present or absent, as there were not different levels of distinction like moderate or strong. However, in later discussions of the ILH, different degrees of search were suggested; moderate search would be a search for the meaning of a given word and strong search would be a search for word forms to express familiar meanings (Laufer, 1999; Nation & Webb, 2011).

Evaluation is another cognitive factor that entails the comparison of an unknown word's form or meaning with other possible words or meanings in order to choose the most suitable one for the context. Evaluation is moderate when a context is provided. Evaluation is strong when learners have to use a word and create a context in which the word fits. There is no Evaluation when learners do not need to decide which words or sense of the word to use. For instance, evaluation is absent when learners read a text with the meanings of unknown words being provided in marginal glosses because the learners do not need to compare the meaning of each unknown word with other words. Evaluation

is moderate when learners read a text containing multiple-choice glosses because the learners need to choose the most suitable meaning for each glossed word that fits the context. Evaluation is strong when learners write sentences using unknown target words because they have to use the target words with other words to create an original context in which the target words fit.

One can calculate an *involvement load* (IL) that represents the estimated effectiveness of an activity. An IL is the total score for an activity. An activity scores 0 points for an absence of a factor, 1 point for a moderate presence of a factor, and 2 points for a strong presence. For example, when an activity involves moderate need (1 point), no search (0 point), and strong evaluation (2 points), the IL of the activity is 3 (1+0+2). The ILH predicts that a task with a higher total score is more effective than a task with a lower total score. Table 1 presents six activities and their ILs. Given that the IL of the writing sentences activity is higher than the one for reading and comprehension questions, the ILH predicts that the former leads to larger vocabulary learning gains.

Table 1: Activities and their Involvement Load Index

Activity	Target word	Need	Search	Evaluation	Involvement load index
Reading and comprehension questions	Glossed in the margin of the text and relevant to the questions	1	0	0	1
Fill-in-the-blanks	Listed with the corresponding L1 translations	1	0	1	2
Writing sentences	Listed with the corresponding L1 translations	1	0	2	3

Writing a composition	Learners chose which words to use by consulting a dictionary	2	1	2	5

1.3 How accurately does the ILH predict the efficacy of activities?

The ILH has been widely discussed by researchers (e.g., Barclay & Schmitt, 2019; Nation & Webb, 2011; Newton, 2020; Schmitt, 2010; Webb & Nation, 2017) and many studies have been conducted to determine whether it accurately predicts the relative efficacy of language activities on vocabulary learning. While research generally provided general support for the ILH prediction (Eckerth & Tavakoli, 2012; Huang, Willson, & Eslami, 2012; Hulstijn & Laufer, 2001; Kim, 2008; Kolaiti & Raikou, 2017; Laufer, 2003), several studies found that their results were not always in line with the prediction of the ILH (Bao, 2015; Folse, 2006; Keating, 2008; Laufer, 2003; Martínez-Fernández, 2008; Rott, 2012; Zou, 2017). For example, Martínez-Fernández (2008) found that activities with higher ILs did not outperform activities with lower ILs. Zou (2017) found that activities with the same IL led to significantly different learning gains. Moreover, in some studies, activities with lower ILs led to greater learning than activities with higher ILs (e.g., Bao, 2015). Because of the inconsistency in the literature, it is difficult to determine the overall validity of the ILH by considering results from individual studies. Systematic and statistical summarization of the ILH studies may produce a summative view of research findings and provide more comprehensive evaluation of the predictive ability of the ILH.

1.4 Potential Approaches to Enhancing the Prediction of the ILH

Because ILH predictions have not always been supported by empirical studies, it might be possible to enhance its predictive ability. Mainly two suggestions have been

made to enhance the ILH prediction. The first is to include more factors that have been reported to be influential on vocabulary learning. For example, Folse (2006) found that an activity with lower IL led to greater vocabulary learning than an activity with higher IL when learners engaged in the former activity repeatedly. This finding suggests that frequency should be included as a factor. Furthermore, Zou (2017) suggested that information organization (i.e., use of chunking and hierarchical organization) should also be included as a factor to enhance the ILH's predictive ability. She found that composition-writing led to larger vocabulary learning gains than sentence-writing and argued that the former involved greater information organization than the latter. This led her to propose distinguishing different type of strong evaluation: sentence level (using a target word in a sentence, e.g., sentence-writing) and composition level (using a set of target words in a composition, e.g., composition-writing).

The second approach to potentially improving the prediction of the ILH is to properly weight each component of the ILH based on its magnitude of influence. When calculating the IL of activities, different components of the ILH (need, search, and evaluation) are assumed to contribute to learning to the same degree. For example, moderate need, present search (search exists as either present or absent), and moderate evaluation are all awarded 1 point for each component and thus assumed to influence learning to the same degree. The same goes for strong need and strong evaluation as both are awarded 2 points. Laufer and Hulstijn (2001) and Kim (2008) mention the possibility that different ILH components might influence vocabulary learning to different degrees. Investigating the degrees of influence of different components on vocabulary learning may indicate the extent to which each factor should be weighted. Meta-analysis of the results of multiple studies testing the ILH may provide a more reliable and summative indication of the extent to which each component should be weighted to more accurately calculate the IL of activities.

1.5 Meta-Analysis

Meta-analysis is a statistical analysis to synthesize the findings of earlier studies (Glass, 1976; Lipsey & Wilson, 2001). Meta-analysis allows researchers to examine (1)

the aggregated effect of certain types of treatments, (2) how consistent the results from earlier studies are, and (3) how characteristics of studies (or treatments) explain the variance of treatment effects.

Many meta-analyses have been conducted to investigate the effectiveness of different types of interventions in the area of Applied Linguistics including studies on the effects of corrective feedback (e.g., Li, 2010; Lyster & Saito, 2010); strategy instruction (e.g., Plonsky, 2011); interaction (e.g., Mackey & Goo, 2007); and using corpus tools (Boulton & Cobb, 2017). To date, two meta-analyses have already been conducted to summarize research findings on L2 vocabulary learning from activities in the classroom (Huang et al., 2012; Won, 2008). Won's (2008) Ph.D. thesis used meta-analysis to investigate the effects of instruction on L2 vocabulary learning. The main findings from the meta-analysis were that (1) most of the instruction was effective and the overall effect size was large (i.e., d = .69), (2) decontextualized learning yielded higher gains than contextualized learning, (3) there were no differences found between studies conducted in EFL and ESL settings, or between instruction with and without provision of L1 supports, and (4) instruction involving multimedia use yielded greater effects than instruction that did not. However, these findings should be considered with caution. Meta-analyses usually examine treatment effects by comparing treatment conditions and control conditions; however, this was not the case in Won's meta-analysis. Won calculated effect sizes by comparing "'special' instruction" or "'innovative' teaching methods" to "traditional instruction" or other comparison groups in each study. Since different studies compared different learning conditions, it is not clear what the effect sizes represent. Furthermore, Won did not clearly state how comparison pairs were selected for those studies including when there were more than two learning conditions. This also makes it difficult to interpret what the effect sizes represent.

Huang, Willson, and Eslami (2012) meta-analyzed 12 studies examining incidental vocabulary learning to investigate the effects of involvement load on learning. They compared output groups (e.g., sentence writing, fill-in-the-blanks, and composition writing) versus non-output groups (e.g., reading activities). The results indicated that (1) output tasks outperformed non-output tasks; (2) results supported the involvement load

hypothesis by revealing that activities with larger involvement indices yielded greater effect sizes, (3) studies with "higher level of design qualities" (e.g., one of the researchers' definitions was studies that controlled participants' prior knowledge of target words by conducting pretests) were more likely to report higher learning gains, compared to studies with "lower level of design qualities" (e.g., one definition was studies that controlled participants' prior knowledge by testing non-participant students with similar or different proficiency levels), (4) time on task had a positive effect on vocabulary learning, (5) reading a combination of expository and narrative texts led to better learning than reading only expository or narrative text, and (6) reading a text with text-target word ratios of less than 2% or equal to 2% led to significantly fewer target words learned compared to reading a text with text-target word ratios of 2%-5%. Although the findings from Huang et al.'s meta-analysis are valuable, they did not comprehensively examine the research literature on ILH.

Won's (2008) and Huang et al.'s (2012) meta-analyses investigating the efficacy of L2 vocabulary learning in the classroom context revealed that students can effectively learn L2 vocabulary by engaging in word-focused activities. Meta-analysis also helps researchers (1) determine the number of studies investigating the issue in question (e.g., Boulton & Cobb, 2017), (2) determine potential biases or methodologies issues, and (3) suggest gaps that future research should fill to deepen the understanding of the area (e.g., de Vos, Schriefers, Nivard, & Lemhöfer, 2018; Shintani, 2015). The present thesis employs a meta-analytic method in order to (1) comprehensively review methodologies and factors of previous studies on the ILH, (2) evaluate the extent to which the ILH accurately predict incidental vocabulary learning gains, (3) examine the relative effects of the components of the ILH, (4) determine how other empirically motivated factors influence learning , (5) update the ILH based on the results of studies that tested the prediction of the ILH, and (6) obtain estimated learning gains for different types of incidental vocabulary learning activities.

1.6 Organization of the Thesis

This thesis adopts an integrated article format and consists of three studies. Each study includes separate introduction, literature review, methodology, results, discussion, and conclusion sections, followed by separate reference lists and appendices. Study 1 (Chapter 2) meta-analyzes earlier studies investigating the ILH to evaluate how accurately the ILH predicts incidental L2 vocabulary learning. Study 1 also examines the relative degree of influence of each component of the ILH (i.e., need, search, evaluation) and examine how other empirically motivated factors (e.g., time on task and frequency) influence learning. Study 2 (Chapter 3) expands on Study 1 and aims to determine whether it is possible to improve the ILH to enhance its accuracy in predicting incidental vocabulary learning. Using the information-theoretic approach, Study 2 identifies an optimal statistical model (i.e., a set of predictor variables) that best predicts vocabulary learning gains reported in the studies testing the ILH. Candidate predictor variables will be selected among the components of ILH and other empirically motivated variables. Based on the resulting statistical model, an IL formula to calculate updated ILs of activities will be created, which predicts the relative effectiveness of incidental vocabulary learning activities more accurately. Study 3 (Chapter 4) systematically reviewed studies that tested the prediction of the ILH to examine their learning conditions and obtain estimated vocabulary learning gains for different incidental vocabulary activities. The learning conditions were categorized into different activity types according to the IL formula's factors, which are identified as useful predictors in Study 2. Using a meta-regression model, the mean learning gains (i.e., percentage of unknown words to be learned) for each activity type will be estimated with their predictive intervals. The final chapter (Chapter 5) summarizes the findings of the three studies in this thesis and discusses the theoretical and pedagogical implications. The chapter also presents the limitations of the three studies and discusses directions for further research.

1.7 References for Introduction and Literature Review

Baddeley, A. D. (1978). The trouble with levels: A reexamination of Craik and Lockhart's framework for memory research. *Psychological Review*, *85*(3), 139–152. http://dx.doi.org/10.1037/0033-295X.85.3.139

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

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348–393. https://doi.org/10.1111/lang.12224

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. http://dx.doi.org/10.1037/0096-3445.104.3.268

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*(2), 227–252. https://doi.org/10.1177/1362168811431377

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology*, *69*(2), 157–169. https://doi.org/10.1111/j.2044-8295.1978.tb01643.x

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.2307/1174772

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

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*(4), 544–557. https://doi.org/10.1111/j.1540-4781.2012.01394.x

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, *12*(3), 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*(2), 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*(3), 406–428. https://doi.org/10.22158/selt.v5n3p406

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*(4), 567–587.

Laufer, B. (1999). Task effect on instructed vocabulary learning: The hypothesis of "involvement." *Selected Papers from AILA'99 Tokyo*, 47–62.

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

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Sage Publications.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA. *Studies in Second Language Acquisition*, *32*(02), 265–302. https://doi.org/10.1017/S0272263109990520

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational Interaction in Second Language Acquisition* (pp. 407–453). Oxford University Press.

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.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt039.0

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

Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(2), 151–171. https://doi.org/10.1016/S0022-5371(77)80044-3

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition*, 1–24. https://doi.org/10.1017/S0272263119000020

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*(4), 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

Rodgers, M. P. H., & Webb, S. (2019). Incidental vocabulary learning through viewing television. *ITL - International Journal of Applied Linguistics*. https://doi.org/10.1075/itl.18034.rod

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.

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

Shintani, N. (2015). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics*, *36*(3), 306–325. https://doi.org/10.1093/applin/amu067

van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*(3), 609–624. https://doi.org/10.1016/j.system.2013.07.012

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 1–17.

Webb, S. (2020). Incidental Vocabulary Learning. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 225–239). Routledge.

Won, M. (2008). *The Effects of Vocabulary Instruction on English Language Learners* [Unpublished doctoral dissertation, Texas Tech University]. https://ttu-ir.tdl.org/ttu-ir/handle/2346/14144

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*(1), 54–75. https://doi.org/10.1177/1362168816652418

Chapter 2

2 To What Extent Does the Involvement Load Hypothesis Predict Incidental L2 Vocabulary Learning? A Meta-Analysis

2.1 Introduction

It is widely accepted that retention of new vocabulary knowledge depends on the amount and quality of attention that individuals pay to the word (Schmitt, 2008; Webb & Nation, 2017). Craik and Lockhart's (1972) *depth of processing* suggested that the greater the degree of semantic and cognitive analysis, the deeper the processing, and the greater the learning. Learning and memory retention improve when new information is used, reformulated, or elaborated. These processes elicit deeper processing by creating connections between pre-existing and new knowledge (Craik & Tulving, 1975). However, one limitation of *depth of processing* was the difficulty in providing a clear and operationalizable definition (e.g., Baddeley, 1978). The theory does not provide a straightforward answer as to whether one learning condition elicits deeper processing over another. It was thus difficult to use *depth of processing* to provide pedagogical suggestions for language learning (Laufer & Hulstijn, 2001).

Laufer and Hulstijn (2001) proposed the *Involvement Load Hypothesis* (ILH) to provide a more operationalizable definition of quality of attention. The ILH postulates that retention of second language (L2) unknown words is conditional upon the *involvement load* (IL) of a task[1], which is determined by one motivational component (*need*) and two cognitive components (*search* and *evaluation*). In response to Laufer and Hulstijn's call, many studies have tested whether the ILH predicts the relative effectiveness of different tasks on vocabulary learning. While some studies report that their results supported the prediction of the ILH (e.g., Eckerth & Tavakoli, 2012; Hulstijn and Laufer, 2001; Kim, 2008), others found that the predictions were not always accurate

[1] In this paper, the terms tasks and activities were used interchangeably to simply refer to language activities.

(e.g., Folse, 2006; Keating, 2008; Rott, 2012). The main reasons for the inconsistency could be due to the weightings of different components (e.g., Kim, 2008; Laufer & Hulstijn, 2001), and other factors such as time on task (e.g., Keating, 2008) and frequency, i.e., how many times participants encountered or used each target word (e.g., Folse, 2006).

The greatest value of the ILH is to provide a tool for language teachers to evaluate the effectiveness of vocabulary learning activities (e.g., Barclay & Schmitt, 2019; Newton, 2020; Webb & Nation, 2017). If the ILH provides accurate predictions of effectiveness, this can help optimize vocabulary learning. Additionally, the ILH is frequently cited to explain the results of empirical studies (e.g., Nguyen & Boers, 2018; Peters et al., 2009). According to Google Scholar, Laufer and Hulstijn (2001) has been cited 1825 times (Aug. 17, 2020). Given the extensive reliance on the ILH in the field, there is a need to systematically synthesize ILH studies to evaluate how well the ILH predicts incidental vocabulary learning. The present study thus adopted a meta-analytic approach to statistically summarize earlier studies assessing the ILH's predictive ability. This study also explored the relative degree of influence of each component (need, search, evaluation) and investigated which factors (e.g., time on task and frequency) moderated incidental L2 vocabulary learning.

2.2 Background

The ILH includes three components: need, search, and evaluation (Laufer & Hulstijn, 2001; see also Hulstijn & Laufer, 2001). By adding the points for the three components, one can calculate an *involvement load* (IL) for each task, which expresses the relative efficacy of the task on vocabulary learning (Laufer & Hulstijn, 2001, p. 16). The need component is the motivational factor referring to whether unknown words are needed to complete a task. Three different levels were suggested for need; need is absent when an unknown word is not required to complete the task (0 points), need is moderate when an external agent (e.g., a task or teacher) asks learners to understand or use the word (1 point), and strong when it is imposed by the learners themselves (2 points). One example of moderate need is when a learner is asked to write a sentence using an

unknown word. Whereas, need is strong when a learner looks up an unknown word in a bilingual dictionary that they want to use in speech or writing.

Search is a cognitive factor and refers to the attempt to find the L2 form of a word or its meaning. Two levels are suggested for search: presence or absence. Search is present when a learner needs to look for a L2 form or its meaning using external resources (e.g., dictionaries or teachers) (1 point). Search is absent when the L2 form and the meaning are provided together in a task (0 points).

Evaluation entails the comparison of an unknown word's L2 form or meaning with other possible words or meanings in order to choose the most suitable one for the context. Evaluation is absent when there is no need to decide which word or sense of the word to use (0 points). Evaluation is moderate when a context is provided (1 point) such as when engaging in a fill-in-the-blanks activity, and the most suitable word for the blanks in a text must be selected from several options. Evaluation is strong when a word must be used in an authentic context. One task that includes strong evaluation is composition writing using target words (2 points).

An *involvement load* (IL) is calculated for each task by adding the points for the three components (Laufer & Hulstijn, 2001, p. 16). For example, a reading activity, where learners are provided with glosses for target words and asked to answer comprehension questions that require learners to understand the target words, involves moderate need (1), no search (0), and no evaluation (0), resulting in a total IL of 1. In contrast, a sentence production activity in which target words and their meanings are provided involves moderate need (1), no search because form and meaning are provided (0), and strong evaluation (2), resulting in a total IL of 3. The ILH predicts that the sentence writing activity leads to greater learning gains than the reading activity as the former has a higher IL.

Two important stipulations of the ILH are that (i) other factors are equal, and (ii) vocabulary learning occurs as incidental learning (as opposed to deliberate learning). First, Laufer and Hulstijn (2001) state that "**Other factors being equal**, words which are processed with higher involvement load will be retained better than words which are

processed with lower involvement load" (p. 15: the emphasis was added by the authors). This means that when other factors (e.g., frequency) are manipulated differently across tasks, learning gains might not be consistent with the predictions of the ILH. Second, the ILH solely focuses on predicting incidental vocabulary learning (as opposed to deliberate vocabulary learning), where learning occurs while engaging in activities without deliberate intention to commit target words to memory. Laufer and Hulstijn (2001, p. 11) argue that, in deliberate vocabulary learning, where students intentionally learn words, learning can be significantly influenced by strategies used by students. Because students may employ different strategies, learning gains may reflect the strategy used by each student instead of the cognitive processes and resources involved in performing the given task. In order to control students' strategy use to ensure that learning gains can be attributed to the features of tasks, the ILH only applies to the realm of incidental vocabulary learning.

2.2.1 Studies testing the Involvement Load Hypothesis[2]

Many studies have tested the ILH by comparing vocabulary learning through different conditions. Research tends to indicate that the relative effectiveness of activities is either completely or partially in line with the prediction of the ILH. Studies supporting the ILH found that tasks with higher ILs resulted in greater learning gains (e.g., Eckerth & Tavakoli, 2012; Tang & Treffers-Daller, 2016) and tasks with the same IL led to similar learning gains (e.g., Kim, 2008; Tang & Treffers-Daller, 2016). Support for the ILH was also provided by Huang, Willson, and Eslami's (2012) meta-analysis. They synthesized 12 studies comparing different learning conditions to evaluate the effect of output tasks (e.g., gap-filling and writing) compared to input tasks (i.e., reading). Results showed that output tasks with a higher IL (i.e., writing) led to greater learning gains compared to output tasks with a lower IL (i.e., gap-filling).

---

[2] In this paper, the term, "test the ILH" was used to refer to testing the prediction made by the ILH and does not refer to testing the ILH as a hypothesis while strictly following the ILH's stipulation that other factors are equal unless explicitly noted so.

However, findings are not always in line with the ILH. Tasks with higher ILs were not necessarily found to lead to greater vocabulary learning than tasks with lower ILs (e.g., Martínez-Fernández, 2008; Yang et al., 2017), and sometimes tasks with lower ILs outperformed those with higher ILs (e.g., Bao, 2015; Wang et al., 2014).

Moreover, sometimes the accuracy of the predictions of the ILH have varied within studies that conducted multiple experiments or administered multiple test formats and/or at multiple time points. Hulstijn and Laufer (2001) found that while an experiment with Hebrew speaking students fully supported the ILH, another experiment with Dutch speaking students provided partial support for the ILH. This points to the possibility that the effect of ILs changes based on the characteristics of participants (e.g., L2 proficiency, the similarities between L1 and L2). Keating (2008) found that the results on an immediate productive recall test offered full support for the ILH as the relative effectiveness of three activities was in line with the predictions of the ILH (i.e., sentence writing led to the greatest gain, followed by gap-filling, then reading with glosses, in that order). However, the same test administered two weeks later only provided partial support for the ILH; no meaningful difference of the mean scores was found between sentence writing and gap-filling. This suggests that the effect of IL might not be observed when the long-term retention of words is examined. Contrasting results have also been reported by many other studies (e.g., Bao, 2015; Rott, 2012; Yang et al., 2017; Wang et al, 2014).

Due to the inconsistent findings, it is difficult to determine the overall validity of the ILH by looking at results from individual studies. Systematically and statistically summarizing the studies that have tested the ILH using meta-regression analysis may produce a summative view of research findings and provide more objective and reliable evaluation of the predictive ability of the ILH.

2.2.2 Relative Contributions of Components of Involvement Load

One reason for the inconsistent findings might be due to the weightings of the different ILH components. The ILH postulates that each component (i.e., need, search, evaluation) contributes to vocabulary learning to the same degree. However, many

researchers, including Laufer and Hulstijn themselves, point out the possibility that the components may have different degrees of influence. Laufer and Hulstijn (2001) state that the search component might have less impact than other components, and Kim (2008) argued that strong evaluation might be the most influential factor for initial learning. A study by Tang and Treffers-Daller (2016) indicated that the influence of evaluation was strongest, followed by need, and that manipulation of the search component did not lead to significantly different learning gains. Synthesizing earlier studies using meta-regression may produce a more robust and summative view of how each component contributes to learning. The results may also allow revisions of the ILH to enhance its accuracy in predicting the potential of tasks for vocabulary learning.

2.2.3 Moderator Variables

There are many variables that may account for the inconsistency in findings between studies investigating the ILH in addition to the components of the ILH (see Appendix A for basic information about the studies).

Time on task. Research has demonstrated that longer tasks tend to be more effective than shorter tasks (e.g., Folse, 2006; Huang et al., 2012; Hulstijn & Laufer, 2001; Keating, 2008). Hulstijn and Laufer (2001) found that a task with a higher IL led to better vocabulary learning than a task with a lower IL, but the former took more time than the latter. Hulstijn and Laufer argue that the superiority of longer tasks may not be due to time on task but to the higher IL because tasks with higher ILs generally take more time than tasks with lower ILs. This led them to suggest treating "time on task as an inherent property of a task, not as a separate variable (p. 549)". However, Keating (2008) found that the ILH did not accurately predict task effectiveness when time on task was controlled, which alludes to the possibility that tasks taking longer lead to greater learning gains regardless of IL. It may be useful to meta-analyze earlier studies to examine which factors, time on task or IL, predict vocabulary learning better, and whether one factor remains influential while controlling the other.

Frequency. Several studies testing the ILH looked at how frequency (i.e., how many times students were exposed to or used each target word) influences the

effectiveness of instructional tasks (e.g., Eckerth & Tavakoli, 2012; Folse, 2006; Y.-T. Lee & Hirsh, 2012). Folse (2006) found that repeating a task with a lower IL three times led to significantly higher scores than completing a task with a higher IL once. This suggests that the number of repetitions might be a more important factor than the ILs of tasks. Eckerth and Tavakoli (2012) investigated the interaction between frequency and IL by examining three tasks at two different frequencies of target word exposures or uses (i.e., once and five times). Their results showed that although both factors clearly contributed to initial word learning, the effect of frequency tended to fade while the effect of IL was more stable for long term retention.

Vocabulary knowledge. Vocabulary knowledge can generally be categorized into three groups: form, meaning, and use (Nation, 2013, p. 49). The majority of studies measured meaning—more specifically, measured form and meaning connections of target words—to test the ILH. Form-meaning connection can be measured either receptively or productively. Receptive tests were either (i) receptive recognition (i.e., select the corresponding L1 translation or L2 synonym for a given word when provided with options) or (ii) receptive recall (i.e., provide the corresponding L1 translation or synonym for a given word). Similarly, productive tests were either (i) productive recognition (i.e., select the corresponding L2 word for a L1 translation or L2 synonym among options) or (ii) productive recall (i.e., provide the corresponding L2 word for a L1 translation or L2 synonym).

Some studies measured vocabulary knowledge related to word use by administering sentence writing tests (e.g., Bao, 2015) and gap-filling tests (e.g., Jahangard, 2013). Sentence writing tests ask participants to use a word in a sentence and assess whether the word is used with semantic and grammatical accuracy. Similarly, gap-filling tests tap into word knowledge in a contextualized format, where participants have to read a sentence and provide a word that fits in the gap with semantic and grammatical accuracy.

Form knowledge has also been measured by using form recognition tests (i.e., Martínez-Fernández, 2008), where participants were asked whether they recognize the appropriate forms of words they encountered during learning.

Another often used approach was to investigate the developmental stage of vocabulary knowledge by using the Vocabulary Knowledge Scale (VKS; Wesche & Paribakht, 1996). VKS captures a word's developmental stage ranging "from complete unfamiliarity, through recognition of the word and some idea of its meaning, to the ability to use the word with grammatical and semantic accuracy in a sentence" (Wesche & Paribakht, 1996, p. 29). VKS may tap into all three aspects of Nation's (2013) framework (i.e., form, meaning, and use).

It is possible that the ILH is fully supported when measuring a certain aspect of vocabulary knowledge, while not supported when measuring another aspect. However, it is difficult to draw a clear conclusion just by looking at individual studies since findings of studies testing the ILH are inconsistent even across studies using the same test formats measuring the same type of word knowledge. For example, while Kim (2008) used the VKS and produced full support for the ILH, Folse (2006) and Zou (2017) used the same VKS format and found that the results did not support the ILH. Furthermore, when studies used multiple test formats, the results tended to be inconsistent. Keating (2008) found that the results on a productive recall test administered immediately after the treatment fully supported the ILH, but those on receptive recall tests only provided partial support. That is, the prediction of the ILH was supported for the comparisons between reading and gap-filling and between reading and sentence writing but the prediction was not supported for the comparison between gap-filling and sentence writing. In contrast, Rott (2012) found that the results on productive recall tests provided full support for the ILH, while those on receptive recall tests only partially supported the ILH; while the differences of learning gains between reading and composition writing and between gap-filling and composition writing were as the ILH predicted, no meaningful difference was found between reading and gap-filling. Given the inconsistency in the literature, it is difficult to draw a clear conclusion about the relationship between the ILH and the aspects of vocabulary knowledge developed. A meta-analytic approach will provide a

summative evaluation on whether the influence of IL varied based on which aspect of vocabulary knowledge was measured by capturing the trend of data through meta-analyzing the results of multiple studies.

Proficiency. Kim (2008) hypothesized that Hulstijn and Laufer's (2001) inconsistent results across different participant groups (i.e., in Israel and the Netherlands) could be due to the participants' L2 proficiencies. Although Kim's results did not support this hypothesis and were in line with the prediction of the ILH regardless of the proficiency of learners, it is possible that learners at different proficiency levels benefit differently from tasks. Since the demands of a task may increase as the IL of the task increase—e.g., reading with glosses has a lower IL and is less demanding than composition writing which has a higher IL—, we hypothesized that participants with a higher proficiency benefit more from tasks with higher ILs, i.e., the effect of IL may be more pronounced for higher proficiency learners than for less proficient learners.

2.2.4 The Current Study

The inconsistency in the results of earlier studies makes it difficult to determine the extent to which the ILH accurately predicts vocabulary learning. These inconsistencies could be because the ILH postulates that all three components of the ILH influence vocabulary learning to the same degree. They might also be due to the many factors such as time on task, frequency, and test format that may influence the effectiveness of tasks.

It is worth recalling that the ILH stipulates that other factors—as opposed to need, search, and evaluation—should be equal. Some of the ILH studies manipulated other factors (e.g., frequency as in Folse, 2006) as well as the IL to investigate whether the prediction of the ILH still holds when the other factors were changed. Although these studies did not test the ILH while strictly following the stipulation of the ILH, it is useful to include these studies and examine (1) the relative usefulness of the empirically motivated factors and (2) whether the effect of ILs changes based on the empirically motivated factors. We adopted a meta-analytic approach to statistically synthesize the studies that strictly manipulated the IL of tasks to investigate how accurately the ILH

predicts vocabulary learning and how different components of the ILH and other empirically motivated factors contribute to incidental vocabulary learning.

Findings of the present study may enhance our understanding of how vocabulary is learned most effectively in incidental contexts by revealing (1) the degree to which each factor included in the ILH contributes to learning, and (2) which other factors need to be addressed in order to obtain a comprehensive model of how learning conditions contribute to incidental vocabulary learning. The findings should also provide pedagogical implications which indicate how language teachers, learners, and material writers can select and design language activities to optimize vocabulary learning.

The study was guided by the following three research questions.

1. To what extent does the ILH predict incidental L2 vocabulary learning?
2. To what extent does each component of the ILH contribute to incidental L2 vocabulary learning?
3. Which empirically motivated factors moderate incidental L2 vocabulary learning in relation to the ILH?

2.3 Method

2.3.1 Literature Search

To answer the research questions, we focused on two types of studies: (a) studies testing the ILH as a hypothesis by manipulating only the ILs of tasks and (b) studies testing whether the relative effects of tasks were predicted by the ILH while manipulating other factors (e.g., frequency, time on task) as well as the ILH components. The latter studies were not testing the ILH as a hypothesis by strictly following the stipulation of the ILH that other factors are equal. To identify studies to include in the meta-analysis, we examined the following electronic databases: Educational Resources Information Centre (ERIC), PsycINFO, Linguistics and Language Behavior Abstract (LLBA),

ProQuest Global Dissertations, Google Scholar, and VARGA.[3] Following earlier suggestions (Oswald & Plonsky, 2010), unpublished research reports such as doctoral dissertations and master's theses were included to comprehensively cover studies of the ILH. We searched for research reports published from 2001 to April 2019 using different combinations of keywords such as involvement load hypothesis, task-induced involvement, involvement load, word/vocabulary, learning/acquisition/retention, and task. Through this electronic database search 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 using Google Scholar to identify the studies that examined vocabulary learning and potentially discussed the ILH. This forward citation search identified 327 more reports. As a result, a total of 1290 research reports were identified and screened according to the following selection criteria.

2.3.2 Inclusion and Exclusion Criteria

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

1. Studies that looked at vocabulary learning from incidental learning conditions were included. We followed Hulstijn's (2001) and Laufer and Hulstijn's (2001) definition of incidental vocabulary learning, where participants were not forewarned about upcoming vocabulary tests before the treatment and participants were not told to commit target words to memory. Studies where participants were told about posttests (i.e., Keating, 2008) and studies where participants were told that the purpose of the study was vocabulary learning (i.e., Maftoon & Haratmeh, 2012) were excluded. Similarly, studies where participants engaged in deliberate vocabulary learning activities (e.g., the keyword technique) were also excluded.

[3] VARGA is an online bibliographical source related to studies on L2 vocabulary acquisition (available at Paul Meara's website: http://www.lognostics.co.uk/varga).

2. Studies testing the ILH and studies that coded the ILH for all learning conditions were included. Studies that mentioned the ILH but did not clearly code learning conditions with the ILH were excluded.
3. Studies reporting enough descriptive statistics to calculate effect sizes (ESs) (i.e., number of participants tested, mean and SD for test scores) were included.
4. Studies that included a learning condition where multiple language tasks were employed, and each task was coded with the ILH were excluded from the analysis. This is because when participants engage in multiple tasks which involve different IL indexes, it is not clear how each component of the ILH contributed to learning gains.
5. We excluded studies reporting research that was already reported in other publications.
6. We excluded studies where activities were not described clearly enough to double-check the authors' coding of the ILH. For example, some studies reported that in certain learning conditions participants had to understand the target words, but did not report how participants might have learned the meanings of target words. Additionally, we excluded studies that failed to describe how learning gains were measured and scored. Because we included non-peer reviewed studies as well as peer-reviewed studies, this criterion also worked as a gate keeper to secure the quality of included studies.

We carefully reviewed the abstracts of the research reports identified through the literature search and retrieved full texts for 137 potential studies (i.e., studies that examined vocabulary learning and mentioned the ILH). We found that 40 studies met all of our criteria. Additionally, we contacted the authors of 14 other studies which were only lacking in the descriptive statistics required for this meta-analysis and gratefully received information from two authors (Hazrat, 2015; Tang & Treffers-Daller, 2016). In all, a total of 42 studies (N = 4628) reporting 398 posttest scores satisfied all of our inclusion and exclusion criteria. These studies comprised 30 journal articles, four master's theses, three book chapters, two doctoral dissertations, two conference

presentations, and one bulletin article. (see Appendix A for basic information about included studies).

2.3.3 Coding

Studies meeting the selection criteria were coded for outcome variables (i.e., descriptive statistics for calculating ESs), IL, moderator variables, and the study identifier (e.g., authors and year).

2.3.4 Involvement Load

We coded the IL of learning conditions in each study. Initially, we planned to follow each studies' coding of the ILH and confirm they matched Laufer and Hulstijn's description for coding tasks. However, we found that some of the coding (10 studies, 23.8%) did not match Laufer and Hulstijn's (2001) description of the ILH. We therefore decided to code IL in two ways: (a) coding conditions following how each author coded, and (b) re-coding conditions strictly following Laufer and Hulstijn's (2001) description for coding (see Appendix B and C for the coding scheme).[4]

To establish the reliability of the ILH coding, we contacted the authors of 8 studies that coded learning conditions differently from our coding and inquired about their rationale for coding and whether they agreed with our coding. We received replies from the authors of four of the studies. Three of them agreed with our new coding and one repeated their explanation from their study. Furthermore, we asked a researcher having expertise in vocabulary research and meta-analysis to double-code the IL for the 10 studies in question by referring to our coding scheme. Intercoder reliability calculated using Cohen's kappa coefficient was $\kappa = .99$, showing it to be high and acceptable. All discrepancies were discussed and resolved.

[4] We also discussed coding with Batia Laufer to solve potential ambiguity in IL coding and to confirm the coding scheme.

2.3.5 Moderator Variables

We coded four moderator variables: time on task, frequency, aspect of vocabulary knowledge, and proficiency. Time on task was calculated by dividing the reported mean (or median) of minutes participants engaged in a task by the number of target words. This was to consider the fact that time on task increases as the number of target words increases in general. Frequency was coded as the number of times that participants encountered or used each target word.[5]

Vocabulary knowledge was coded as either form (word form recognition: asking to select the appropriate forms of the words encountered in a text, Martínez-Fernández, 2008), meaning (i.e., form-meaning connection: receptive recall/recognition and productive recall/recognition), use (i.e., sentence writing and gap-filling), and the VKS. Following previous meta-analyses of vocabulary learning (e.g., de Vos, 2018; Uchihara et al., 2019), we further divided form-meaning connection into two categories based on its sensitivity: recall and recognition. When the learning gain was measured with Wesche and Paribakht's developmental scale, the VKS, we assigned a separate category since this test taps into all three aspects: form, meaning, and use.

Following earlier meta-analyses (e.g., Boulton & Cobb, 2017), participants' L2 proficiency was coded as (a) beginner, (b) intermediate, or (c) advanced based on the reported proficiency. [6]

---

[5] Frequency of the repetition in encountering/using "the same word" can be operationalized in different ways (Reynolds & Wible, 2014) such as the word type, lemma, flemma, and word family. Unfortunately, none of the studies provided a clear explanation of how the repetition of "the same word" was operationalized.

[6] Earlier meta-analyses tend to note the difficulty in analyzing participants' L2 proficiency (e.g., Boulton & Cobb, 2017; Jeon & Yamashita, 2014; J. Lee, Jang, & Plonsky, 2015). Although 23 studies out of 42 included studies (45.2%) reported participants' L2 proficiencies, their judgements were based on various criteria. Six studies (26.1%) judged proficiency based on the level of the classes or schools that participants belong to (e.g., Martínez-Fernández, 2008; Tang & Treffers-Daller, 2016; Yang, Shintani, Li, & Zhang, 2017), 3 (13.0%) referred to the results of standardized English proficiency tests (e.g., Oxford Placement Test, Jahangiri & Abilipour, 2014), 2 (8.7%) administered national English tests or entrance examinations and used the results to judge proficiencies (e.g., Zou, 2017), and 1 (4.3%) referred to the results of a

Coding procedure. Following earlier meta-analyses and suggestions (Plonsky & Oswald, 2015), four researchers in the field of Applied Linguistics were included in the coding process. First, two researchers, 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 consistent. Finally, one author carefully coded the 42 studies, then randomly selected 22 studies (52.4%) which were then separately double-coded by two other researchers in the field of Applied Linguistics who had carried out previous meta-analytic studies. We calculated the inter-coder reliabilities using Cohen's Kappa κ and found the agreement rate was high and acceptable at κ = .99 and .98 for each double-coder. All discrepancies were discussed and resolved. All data (completed coding sheet) are publicly accessible via the Open Science Framework.

2.3.6 Data Analysis

In this meta-analysis, we dealt with studies examining vocabulary learning in multiple learning conditions. We used multilevel meta-regression analysis (Cheung, 2014; H. Lee et al., 2018) to account for different sources of variance: variance between studies and within studies as well as sampling variance.[7] Many studies reported posttest scores that were dependent due to a sampling error (e.g., the same participants were tested repeatedly or with different test formats), which potentially causes a Type I error inflation. To deal with this, we applied the cluster-robust variance estimation (Hedges et al., 2010) with small sample adjustments (Tipton, 2015; Tipton & Pustejovsky, 2015).

---

vocabulary levels test (i.e., Beal, 2007). The rest (11 studies, 47.8%) did not report how they determined the proficiency level.

7 Three-level meta-regression models used in the current study can be seen as an extension of the traditional random-effects model (Hedges & Olkin, 1985), which is equivalent to a two-level meta-regression model (e.g., Fernández-Castilla et al., 2020).

Effect size calculation. Following earlier meta-analyses on vocabulary research (Swanborn & de Glopper, 1999; Yanagisawa et al., 2020), we calculated relative learning gain.

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

Each calculated ES was weighted using the sampling variance of the posttests scores (see Appendix D for detailed ES and sampling variance calculation formulas) (see also Card, 2012; Hox, 2010). By using this relative learning gain as ES and multilevel meta-regression, posttest scores across different studies were comparable while variance between- and within-studies was accounted for.

Analysis procedure. All of the analyses were conducted in the R statistical environment (R Core Team, 2017) using the metafor package (Viechtbauer, 2010) and the clubSandwich package (Pustejovsky, 2018). Three level meta-regression models (Cheung, 2014; H. Lee et al., 2018) were used to model three different sources of variance, i.e., 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). The ESs of immediate and delayed posttest scores were analyzed separately. The significance level was set at 5%. P-values lower than .10 were also interpreted as indicating that there was a trend effect and the effect of the factor in question was investigated further by examining the size and direction of the coefficient and its CI.

To answer the first research question, we conducted a three-level meta-regression fitting a statistical model where the IL predicts ESs. Using equations in Cheung (2014), we calculated explained variance at different levels, within- and between-study levels. Explained variance at the within-study level indicates the proportion of explained variance in ESs within the same study. This corresponds to the variance explained by the variables while the effects of the characteristics of target words and participants are held constant. The explained variance at the between-study level and the overall explained variance (both at within- and between-study levels) were also calculated to examine the

explanatory power of the ILH across studies. Because these explained variances are non-negative by definition, negative values were truncated and interpreted as zero (Cheung, 2014).

To answer the second research question about the degree to which each component of the ILH contributes to learning, we inserted each component as a predictor variable into statistical models. Finally, to answer the third research question, each moderator variable was inserted into the model, firstly as a main effect with and without controlling the effect of IL, and secondly as a main effect and interaction with IL. Examining the main effect reveals how the moderator variable contributes to learning independently from IL. Examining the interaction effect reveals how the influence of IL was moderated by the moderator variables (also, see Appendix E for publication bias analysis and additional analyses).

2.4 Results

2.4.1 Research Question 1: To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning?

Among the 398 ESs included in the analysis, 20 ESs (5.0%) were from learning conditions where the IL was coded as 0, 76 ESs (19.1%) had an IL coded as 1, 139 ESs (34.9%) were coded as 2, 137 ESs (34.4%) were coded as 3, and 26 ESs (6.5%) had an IL coded as 4. IL can theoretically go up to 5 (strong need, search, and strong evaluation); however, none of the studies included a learning condition where the IL was 5. The intra-class correlations (ICCs; Cheung, 2014) calculated based on intercept only models showed that for immediate posttests, 50.8% of the variance was due to between-study variance, and 49.2% of the variance was due to within-study variance. Similarly, for delayed posttests, 62.2% of variance was due to between-study variance, and 37.8% was due to within-study variance. This indicates that variation in learning gains was more likely to be due to external factors (e.g., participants, target items, learning contexts) compared to internal factors (e.g., learning conditions) and this trend was more pronounced for learning retention measured with delayed posttests.

To answer the first research question, we conducted a multilevel meta-regression analysis (See Table 1 for the overall results). The analysis revealed that ESs were significantly predicted by IL on both immediate and delayed posttests ($p < .001$ for both). For ESs on immediate posttests, IL explained 1.3% of between-study level variance and 29.1% of the within-study level variance. For ESs on delayed posttests, IL explained 0% of between-study level variance and 26.5% of within-study level variance. The overall explained variance was 15.0% for immediate posttests and 5.1% for delayed posttests.

Table 1: Results of the Extent to Which the ILH Predicts Incidental L2 Vocabulary Learning

	Immediate Posttests		Delayed Posttests	
	b [CI]	*p*	*b* [CI]	*p*
Intercept	.230 [.144, .315]	< .001	.174 [.088, .260]	< .001
IL	.090 [.056, .125]	< .001	.070 [.041, .098]	< .001
Total R^2	.150		.051	
Between-study R^2	.013		-.079	
Within-study R^2	.291		.265	
k	37		34	
n	207		191	

Notes. IL = involvement load. CI = 95% confidence interval. *k* = number of studies. *n* = number of ESs.

2.4.2 Research Question 2: To what extent does each component of the involvement load hypothesis contribute to incidental L2 vocabulary learning?

The studies included in this meta-analysis had a variety of learning activities (e.g., reading, listening, matching, retelling, and writing with a dictionary) with different combinations of each component of the ILH (see Appendix F for the number of ESs for each combination of components of the ILH and examples of activities). The combination least frequently investigated was a condition involving moderate need, search, and no evaluation (12 ESs, 3%), and the most frequently investigated combination was a condition involving moderate need, no search, and moderate evaluation (128 ESs, 31.9%). The most frequently used learning tasks were reading (122 ESs, 30.6%), followed by writing (111 ESs, 27.9%) and fill-in-the-blanks (105 ESs, 26.4%), and these three tasks accounted for the majority of learning conditions (84.9%).

We administered a series of meta-regression analyses to examine how different components of the ILH (i.e., need, search, and evaluation) were related to ESs (see Table 2 and Table 3 for the results on immediate and delayed posttests, respectively). The results of immediate posttests showed that when each component of the ILH was examined as a predictor variable, need and evaluation were significantly predictive of learning gains: $b = .301$, $p = .007$ for moderate need, $b = .135$, $p < .001$ for moderate evaluation, and $b = .223$, $p < .001$ for strong evaluation. These results indicate that when learning conditions include a need component, ESs were estimated to be 30.2% higher compared to learning conditions that did not include need. With conditions involving moderate and strong evaluation components, ESs were estimated to be 13.9% and 22.6% higher, respectively, compared to learning conditions that did not include either evaluation component. In contrast, the search component was not a significant predictor of ESs ($p = .515$). The total explained variance and the explained variance at the within-study level was the greatest when evaluation was used as a predictor (14.8% for total and 29.9% for the within-study level), followed by need (4.2%, 16.6%), then search (0.9%, 0%). The explained variance at the between-study level was relatively small for all

components (0.2% for evaluation, 0% for need, and 1.9% for search). This small explained variance may be due to the fact that the ILH components were related to learning conditions that differed within each study and factors that were different across studies (e.g., characteristics of participants and target words) were not considered here.

To determine an estimated degree of contribution of each component while controlling the influence of the other components, we conducted a multiple meta-regression in which three components were included as predictors altogether. The results again showed that moderate need, moderate evaluation, and strong evaluation were significantly predictive of ESs ($p = .034$, $p < .001$, $p < .001$, respectively), while the search component was not a significant predictor ($p = .344$). The intercept ($b = .149$) indicates that learning conditions involving no need, no search, and no evaluation led to 15.4% of unknown words learned. The learning gains increased by 20.0% when learning conditions included only need, showing the importance of a need component even when search and evaluation are not present. For the evaluation component, the effect of strong evaluation ($b = .191$) was almost twice the effect of moderate evaluation ($b = .103$). Estimated learning gains increased by 30.3% when learning conditions included moderate evaluation and by 39.1% when including strong evaluation compared to when need and search were not present. Additionally, the difference between moderate evaluation and strong evaluation was significant ($b = .088$, 95% CI [0.470, 0.128], $p < .001$). The search component was not significantly predictive of ESs and the direction of influence was negative ($b = -.023$, 95% CI [-.073, .028]).

Table 2: Results of the Extent to Which the ILH Predicts Incidental L2 Vocabulary Learning on Immediate Posttests

	Immediate Posttests							
	Need		Search		Evaluation		Multiple Regression	
	b [CI]	*p*	b [CI]	*p*	b [CI]	*p*	b [CI]	*p*
Intercept	.155 [-.003, .313]	.053	.433 [.376, .502]	< .001	.307 [.241, .374]	< .001	.149 [-.012, .313]	.065
Moderate Need	.301 [.124, .478]	.007					.200 [.020, .380]	.034
Search			.020 [-.045, .086]	.515			-.023 [-.073, .028]	.344
Moderate Evaluation					.135 [.077, .194]	< .001	.103 [.057, .148]	< .001
Strong Evaluation					.223 [.149, .298]	< .001	.191 [.131, .250]	< .001
Total R^2	.042		.009		.148		.139	

Between-study R^2	-.078	.019	.002	-.084
Within-study R^2	.166	-.002	.299	.369

Notes. IL = involvement load. CI = 95% confidence interval. *k* = number of studies.

The analysis of the delayed posttests revealed similar results (see Table 3). When each component of the ILH was examined as a predictor variable, need and evaluation were significantly predictive of learning retention: $b = .195$, $p = .017$ for moderate need, $b = .111$, $p < .001$ for moderate evaluation, and $b = .179$, $p < .001$ for strong evaluation. In contrast, search was not significant ($p = .490$). The total explained variance and the explained variance at within-study-level was greatest when evaluation was used as a predictor (9.4% and 29.2%, respectively), followed by need (1.9%, 10.5%), then search (0.5%, 0.1%).

The analyses of the multiple meta-regression for the delayed posttests including all three components to predict learning retention revealed the same trend as with the immediate posttests; moderate need, moderate evaluation, and strong evaluation were significantly predictive of ESs ($p = .043$, $p < .001$, $p < .001$, respectively), while the search component was not a significant predictor ($p = .197$). The intercept ($b = .127$) indicated that learning conditions involving no need, no search, and no evaluation led to 12.7% of unknown words retained. Including moderate need increased retention by 12.6%, leading to total learning gains of 25.3% of unknown words learned. Similarly, both moderate evaluation and strong evaluation components were significantly predictive of ESs ($b = .094$, $p = .001$ and $b = .156$, $p < .001$, respectively). Learning retention increased by 21.7% when learning conditions included moderate evaluation and by 27.9% when including strong evaluation compared to when need and search were not present. The difference between moderate evaluation and strong evaluation was significant ($b = .064$, 95% CI [.027, .101], $p = .002$). The search component was not significantly predictive of ESs and the direction of the influence was negative ($b = -.049$, 95% CI [-.129, .030], $p = .197$). We carried out additional analyses regarding the search component and confirmed that the influence of search was not due to its different operationalizations (see Appendix E).

Table 3: Results of the Extent to Which the ILH Predicts Incidental L2 Vocabulary Learning on Delayed Posttests

	Delayed Posttests							
	Need		Search		Evaluation		Multiple Regression	
	b [CI]	*p*	b [CI]	*p*	b [CI]	*p*	b [CI]	*p*
Intercept	.142 [.006, .279]	.043	.331 [.264, .398]	< .001	.224 [.154, .293]	< .001	.127 [.000, .254]	.050
Moderate Need	.195 [.055, .336]	.017					.126 [.006, .247]	.043
Search			-.020 [-.084, .043]	.490			-.049 [-.129, .030]	.197
Moderate Evaluation					.111 [.063, .159]	< .001	.095 [.050, .141]	.001
Strong Evaluation					.179 [.128, .231]	< .001	.160 [.115, .205]	< .001
Total R^2	.019		.005		.094		.104	
Between-study R^2	-.034		.008		-.027		-.034	
Within-study R^2	.105		.001		.292		.330	

Notes. IL = involvement load. CI = 95% confidence interval.

2.4.3 Research Question 3: Which empirically motivated factors moderate incidental L2 vocabulary learning in relation to the involvement load hypothesis?

A series of multiple-regression analyses were carried out for each moderator variable. The main effect indicates the influence of each moderator variable on learning gains. The main effect while controlling the influence of IL indicates whether the effect of the variable remained even when the influence of IL was controlled. The interaction indicates whether the variable moderated the effect of IL on vocabulary learning. The results of these analyses are presented in Table 4.

Time on task. Twenty-six of the 42 studies (61.9%) reported how long (minutes) participants engaged in learning conditions. The mean minutes per word was 3.13 (*SD* = 1.83, Median = 3, Mix = 0.21, Man = 7.50). The analyses revealed that the main effects of time on task was significant both on immediate (p = .041) and delayed posttests (p = .007). This indicates that learning conditions that take longer yield larger learning gains than those that take less time. However, when IL was controlled, the main effects were not significant on both immediate (p = .577) and delayed posttests (p =.266) while IL stayed as a significant predictor. This suggests that longer learning conditions do not necessarily lead to greater learning gains, but rather learning conditions with larger ILs tend to take longer, and IL contributes to learning more than time on task. This was confirmed by a three-level meta-regression without weighting to examine the relationship between IL and time on task. These results showed that IL and time-per-word were significantly correlated (standardized b = .353, 95% CI [.164, .542], p = .001). There were no significant interactions on immediate or delayed posttests (p = .168, p = .208, respectively) indicating that time on task does not moderate the effect of IL.

Frequency. The majority of the ESs (327, 82.2%) were from learning conditions where participants encountered or used each target word only once, and a relatively small number of ESs were from conditions involving multiple encounters or uses of target

words: 12 ESs (3%) for two times, 11 ESs (2.8%) for three times, 48 ESs (12.1%) for four times.

The analyses of immediate posttests showed that there was a trend of the main effect (b = .045, 95% CI [-.005, .122], p = .064), showing that frequency was positively correlated with the learning gain. This main effect was more clearly pronounced when IL was controlled (b = .083, 95% CI [.019, .147], p = .021). This suggests a trend whereby encountering or using the same target words multiple times increases learning gains. The estimated learning gain increased by 8.3% as frequency increased by 1 when controlling the effect of IL. In contrast, the effect of frequency disappeared when looking at the ESs of delayed posttests (p = .856 for when IL was not controlled and p = .898 for when IL was controlled). The interaction between frequency and IL was not significant for immediate posttests (p =.497) or delayed posttests (p = .526). This suggests that the effect of IL did not change greatly regardless of frequency.

Aspect of vocabulary knowledge. The most frequently administered test format was receptive recall (28 studies) followed by VKS (11), productive recall (7), use (4), receptive recognition (3), and form recognition (1). No other test formats were used among the included studies—none of the included studies measured productive recognition (meaning cue). These test formats were categorized into five groups based on the aspect of vocabulary knowledge: (i) form (form recognition; 12, 3%), (ii) form-meaning: recall (receptive and productive recall; 256 ESs, 64.3%), (iii) form-meaning: recognition (receptive recognition; 26, 6.5%), (iv) use (sentence writing and gap-filling; 25 ESs, 6.3%), and (vi) the VKS (79 ESs, 19.9%).

Since a Wald-test with small sample adjustments sometime did not calculate p-values (probably due to the great degree of imbalance of sample sizes across different test formats, especially form recognition measured by only one study), Wald-tests without small sample adjustment were carried out throughout for this moderator variable of vocabulary knowledge for the sake of consistency. To test estimated coefficients of meta-regression, the cluster-robust variance estimation with small sample adjustments was used.

The analyses of the Wald-test on immediate posttests found significant main effects with and without controlling IL ($p < .001$, $p < .001$, respectively). While controlling the influence of IL, learning gains were the highest when measured for form, followed by form-meaning recognition, form-meaning recall, VKS, and use, in that order. Subsequent multiple comparisons while controlling IL showed that form led to significantly higher learning gains than form-meaning recall ($p = .013$), VKS ($p < .001$), and use ($p < .001$). Form-meaning recognition was higher than form-meaning recall ($p < .100$), VKS ($p = .012$), and use ($p = .012$). Form-meaning recall was significantly higher than use ($p = .013$). No significant difference was found across the other comparisons: form vs. form-meaning recognition ($p = .537$) and form-meaning recall vs. VKS ($p = .364$).

The analyses of the main effects on delayed posttests produced similar results, in that vocabulary knowledge was significant with and without controlling IL ($p < .001$, $p < .001$, respectively). While controlling the IL, learning gains were the highest when measured for form, followed by form-meaning recognition, form-meaning recall, use, and VKS, in that order. Form led to higher learning gains than form-meaning recall ($p = .096$), use ($p = .041$), and VKS ($p = .007$). Form-meaning recognition had higher learning gains than form-meaning recall ($p = .041$), use ($p = .003$), and VKS ($p = .001$). Form-meaning recall led to significantly higher learning gains than use ($p = .030$), and VKS ($p = .033$). There were no clear differences between form and form-meaning recognition ($p = .715$) and between use and VKS ($p = .235$).

The analyses of Wald-tests on an interaction between vocabulary knowledge and IL did not reach statistical significance on immediate posttests ($p = .112$) but reached significance on delayed posttests ($p = .011$). On immediate posttests, although the Wald-test did not reach significance, it is useful to examine the trend of the effect (e.g., Plonsky, 2015). The influence of IL was the most pronounced on form-meaning recall, followed by use, form-meaning recognition, form, and VKS, in that order. The coefficients of meta-regression analyses revealed that the influence of IL was stronger on form-meaning recall compared to those on VKS ($p = .034$) or form ($p = .059$). There was also a trend that the effect of IL was more pronounced on use than VKS ($p = .066$).

On delayed posttests, the influence of IL was the most pronounced on form, followed by form-meaning recognition, form-meaning recall, use, and VKS in that order. The influence of IL was more pronounced on knowledge of form compared to use (p = .045) or VKS (p = .017). However, this has to be interpreted with caution since only one study (i.e., Martínez-Fernández, 2008) accounted for form. These results suggest that IL had weaker effects on the development of use knowledge or VKS's developmental stages of word knowledge compared to the development of form and form-meaning knowledge.

Proficiency. Out of 42 studies, 23 (54.8%) reported participants' L2 proficiency: 4 studies (16.7%) recruited beginners, 15 studies (62.5%) included intermediate learners, and 5 studies (20.8%) involved advanced learners. The analyses of a Wald-test on immediate and delayed posttests did not find any main effects (p = .988, p =.746, respectively), main effect while controlling IL-index (p =.881, p =.613), or interactions (p = .275, p = .652). These results indicate that in contrast to our hypothesis, there was no clear advantage of higher proficiency learners over less proficient learners for tasks with higher ILs.

Table 4: Results of the Moderator Analyses on Immediate Posttests and Delayed Posttests

			Main Effect		Main effect while IL controlled		Interaction	
Variable	*k*	*n*	*b* [CI]	*p*	*b* [CI]	*p*	*b* [CI]	*p*
1. Task Variables								
(1) Time on Task								
Immediate	24	146	.065 [.004, .126]	.041	.016 [-.041, .075]	.577	-.016 [-.042, .009]	.168
Delayed	23	131	.056 [.022, .090]	.007	.018 [-.016, .053]	.266	-.012 [-.033, .008]	.208
(2) Frequency								
Immediate	37	207	.045 [-.005, .122]	.064	.083 [.019, .147]	.021	-.008 [-.070, .054]	.497

Delayed	34	191	-.008 [-.141, .126]	.856	.005 [-.108, .118]	.898	-.010 [-.059, .039]	.526
2. Methodological Variables								
(1) Aspect of Vocabulary Knowledge								
Immediate								
Form-meaning recall	25	123	-ref.-		-ref.-		-ref.-	
Form-meaning recognition	3	13	.256 [-.291, .803]	.091	.276 [-.185, .738]	.100	-.052 [-.280, .176]	.371
Form	1	6	.320 [.195, .445]	.008	.337 [.206, .467]	.013	-.066 [-.135, .004]	.059

Use	4	15	-.105 [-.161, -.050]	.007	-.101 [-.150, -.051]	.010	-.024 [-.091, .043]	.366
VKS	10	50	-.021 [-.171, .129]	.767	-.063 [-.207, .080]	.364	-.072 [-.136, -.007]	.034
Delayed								
Form-meaning recall	26	133	-ref.-		-ref.-		-ref.-	
Form-meaning recognition	3	13	.260 [-.028, .547]	.058	.276 [.037, .515]	.041	.006 [-.154, .166]	.852
Form	1	6	.291 [-.069, .651]	.068	.310 [-.193, .813]	.096	.087 [-.049, .223]	.105
Use	3	10	-.089 [-.153, -.024]	.032	-.084 [-.141, -.028]	.030	-.014 [-.107, .080]	.617

VKS	7	29	-.111 [-.240, .017]	.073	-.130 [-.241, -.020]	.033	-.034 [-.099, .032]	.245
3. Learner Variable								
(1) Proficiency								
Immediate								
Beginner	4	18	-ref.-		-ref.-		-ref.-	
Intermediate	12	78	-.010 [-.315, .294]	.934	-.060 [-.341, .221]	.603	-.020 [-.094, .054]	.487
Advanced	5	20	.015 [-.344, .375]	.920	-.007 [-.329, .314]	.956	.040 [-.049, .130]	.294
Delayed								
Beginner	4	18	-ref.-		-ref.-		-ref.-	

Intermediate	13	87	-.051 [-.303, .202]	.624	-.083 [-.308, .143]	.385	-.021 [-.087, .046]	.417
Advanced	5	23	-.106 [-.411, .199]	.426	-.113 [-.408, .183]	.385	-.028 [-.114, .058]	.435

Notes. *k* = number of studies. *n* = number of ESs. *b* = estimated unstandardized coefficient. -ref.- = reference level. CI = 95% confidence interval. *p* = *p*-value for a significant test for the coefficient. Interaction = coefficient for the interaction effect between the moderator variable and Involvement Load (IL). VKS = Vocabulary Knowledge Scale tests.

2.5 Discussion

In answer to the first research question, the findings provided moderate support for the ILH. The results found a clear correlation between IL and relative vocabulary learning gains on both immediate and delayed posttests. This indicates that learning gains tend to increase as the IL of a task increases and suggests that the ILH is a useful framework that adequately explains the relationship between learning conditions and learning gains.

On the other hand, the results suggest that the predictive ability of the ILH is not very high. The explained variance indicated that the ILH explained 29.1% and 26.5% of the variance at a within-study level on immediate and delayed posttests, respectively. This suggests that only about one third of the variance in incidental vocabulary learning and retention can be accurately predicted by the ILH even when the influence of target words and participant characteristics are controlled. Moreover, the ILH explained 15.4% and 5.5% of overall variance (i.e., variance at both within- and between-study levels) on immediate and delayed posttests, respectively. This means that the ILH may not provide accurate estimations of learning gains across studies, where different participants and target words were utilized to test the ILH. The explained variance at the within-studies level provides a more accurate measure of the predictive power of the ILH because studies that test the ILH tend to meet the stipulations of the ILH. In contrast, the overall explained variance does not provide an accurate measure of the predictive power of the ILH because the second stipulation of the ILH, that other factors are equal, is not met. However, the overall explained variance provides a useful indication of how other factors may affect incidental learning.

The low predictive ability of the ILH could be due to the fact that the ILH treats its different components as contributing to learning to the same degree. It might also be due to other factors impacting learning beyond the ILH. The results showed that none to very little variance was explained by the ILH (1.3% on immediate and 0% on delayed posttests) at the between-study level. This should be expected because the ILH predicts

the relative efficacy of tasks by including factors related to learning conditions and does not include other factors (e.g., characteristics of participants and target words) that vary across studies. However, these results highlight the fact that learning gains greatly differ across studies and suggest that considering other factors that vary across studies (e.g., characteristics of students [e.g., vocabulary size] and target words [e.g., number of letters, similarities to L1]) may enhance the prediction of learning gains.

Another potential reason for the relatively low explained variance was that several studies did not strictly follow the ILH's stipulation that other factors are equal. For example, four studies included tasks where frequency was not the same across activities (i.e., Ansarin & Bayazidi, 2016; Folse, 2006; Jahangiri & Abilipour, 2014; Lee & Hirsh, 2012). These studies tend to indicate that frequency should be considered when determining the IL of tasks. Although there is value in examining how other factors contribute to incidental vocabulary learning, the inclusion of studies that do not strictly adhere to the ILH stipulations may impact the explained variance. Therefore, to determine how the inclusion of these studies affected the prediction of the ILH, we reran the analysis while only including the studies where frequency was equal across tasks. The results showed that the variance explained by the ILH increased slightly at all three levels on immediate posttests (by 1.8% for the total explained variance, by 1.5% at the between-study level, and by 2.9% at the within-study level). The total explained variance increased on delayed posttests by 0.5%, while the explained variance at the other levels remained the same. Although authors of these studies tend to argue that frequency was a more important factor than the IL of the task, they may have underestimated the effect of IL because they did not keep other factors constant.

Furthermore, to control for the effect of time on task, some studies provided the same amount of time for participants to complete activities across different tasks. However, this approach might have provided participants with too much or not enough time to complete the task because time on task changes based on the characteristics of a task. To examine whether such manipulation of time on task influenced learning gains and the effect of IL, we carried out another sensitivity analysis. In 10 studies (23.8%), participants were provided with the same time on task across different learning conditions

(Cheng, 2011; Hirata & Mori, 2008; Hyun, 2011; Jahangiri & Abilipour, 2014; Keyvanfar & Badraghi, 2011; Kim, 2008; Konno et al., 2009; Tang & Treffers-Daller, 2016; Tsubaki, 2012; Yang et al., 2017). We carried out meta-regression analyses with an indicator variable specifying whether or not participants were provided with the same time to complete a task across different tasks as well as the IL and the interaction between the indicator variable and the IL. The meta-regression analyses were conducted separately for immediate and delayed posttests. The results indicated that equal time on task did not significantly influence the ESs (b = .026, p = .682 on immediate and b = .001, p = .983 delayed posttests) or the effect of IL (b = -.036, p = .204 on immediate and b = -.017, p = .433 delayed posttests), although the effect of IL was slightly less pronounced. This suggests that although there is a small chance that the effect of IL is less pronounced when time on task is equal across different tasks, we could not find clear evidence of that.

This meta-analysis also revealed some inconsistency in IL coding of conditions across studies. Eleven studies (26.2%) coded their learning conditions differently from Laufer and Hulstijn's (2001) description of the ILH (see the completed coding scheme that is publicly available online). On the one hand, this inconsistency may be due to researchers' different understandings of the ILH. On the other hand, this might also reflect the difficulty in quantifying factors related to learning conditions. That is, the ILH components might sometimes be difficult to code dichotomously. For example, Hulstijn and Laufer (2001) operationalized no search as the prevision of marginal glosses as proposed in Laufer and Hulstijn (2001, p. 15). However, learners still search for information about the meaning of the words by directing their attention from reading to the marginal glosses, thus one might wonder "how far" learners have to search to claim that the task includes a search component. Whether search is present when learners' have to search for a word in a glossary inserted at the end of the text (as in Tang and Treffers-Daller, 2016), or learners need to use a dictionary (as described as one example in Laufer & Hulstijn, 2001) is not clear. The same goes for the need component; learners might be internally motivated to use target words even when the task was assigned by the teacher. Creating clearer criteria explaining how different conditions should be coded may enhance the consistency of coding across studies and enable the reliable evaluation of

different conditions. Future studies proposing or revising a hypothesis regarding L2 learning are encouraged to provide different examples of how different learning tasks should be coded to assist researchers who aim to test the hypothesis.

2.5.1 Relative Effects of each Component of the Involvement Load Hypothesis

In answer to the second research question, the analysis indicated that each component contributes differently to learning. The evaluation and need components significantly contributed to relative vocabulary learning gains, while the search component did not.[8] The evaluation component alone explained the largest proportion of the overall variance in learning gains (14.8% for immediate posttests, 9.4% for delayed posttests), while need explained quite a small proportion of variance (4.2%, 1.9%). Additionally, strong evaluation led to greater learning gains than moderate evaluation.

Laufer and Hulstijn (2001) primarily based the need component of the ILH on Gardner and Lambert's (1972) two categorizations of motivation: integrative (i.e., generated by learners themselves, corresponding to the ILH's strong need) and instrumental (i.e., generated by the instructional orientations, corresponding to moderate need). The results showing the significant contribution of moderate need suggest that instructional manipulation increases learners' motivation to learn target words (Laufer & Hulstijn, 2001). Given that strong need has not been investigated in studies of the ILH, how tasks generating integrative motivation compare with those generating instrumental motivation remains to be determined. Additionally, it may be useful for future studies to

[8] One might wonder whether the lack of contribution of search to learning could have been due to the fact that the majority of the studies included learning conditions without search (see Appendix F). However, although the number of ESs from conditions with search was smaller than those from conditions without search, it was quite large (82 ESs), so the lack of a search effect may not be due to the shortage of data examining the search component. Furthermore, the direction of the estimated coefficients for search while controlling other factors was negative, suggesting that adding further data from learning conditions with the search component may not reveal the positive influence of search—at least a strong effect seems unlikely to be observed. Additional analyses confirmed that the different operationalizations of search may have little influence on the results (see Appendix E).

expand upon the ILH's motivational components to reflect more recent research such as Deci and Ryan's (1985) *self-determination theory* (see also, Noels, Pelletier, Clément, & Vallerand, 2000, for its application to L2 learning).

The results showing the significant contribution of evaluation suggest that an elaborative process in which learners compare a word with other words, contrast a specific meaning of a word with other meanings, or combine a word with other words to create sentences facilitates the learning of the new words. The advantage of strong evaluation over moderate evaluation was observed, and this advantage may be explained by the fact that tasks involving strong evaluation require learners to pay attention not only to the form-meaning connection of a word but also to syntagmatic and collocational knowledge of the word (Laufer & Hulstijn, 2001, p. 15; see also Kaivanpanah and Miri (2018).

The findings are important because they show that all three components should not be considered equally influential. The results indicated that evaluation, especially strong evaluation was the component that contributed to learning most. This highlights the value of productive activities such as writing and speaking where learners use target words in original sentences or compositions. Although a significant effect of need was found, need alone explained only a small proportion of variance in learning gains. On the one hand, this suggests that only looking at whether an activity involves need does not provide a useful prediction of vocabulary learning. On the other hand, given that need was still significant when the influence of evaluation and search was controlled, the findings reveal that need positively influences learning even when evaluation or search is absent. Therefore, educators and material writers should be encouraged to use activities where students feel motivated to understand and/or use unknown target words. Given the fact that the effect of search was not found, the findings suggest that it is more effective for students to use target words productively in sentences or compositions with the provision of the meanings of unknown words in a list or glosses instead of spending time looking up the words in a dictionary.

At first glance, the non-significant search effect might appear to contradict studies showing the benefits of dictionary use (Cho & Krashen, 1994; Knight, 1994) and the effect of look-up frequency (Hill & Laufer, 2003; Peters, 2007).[9] This could be explained by the differences between these studies and studies testing the ILH. In studies of dictionary use participants were provided with dictionaries as the only source of obtaining information about unknown target words. If participants did not consult a dictionary, they could not establish form-meaning mappings unless they successfully guessed the meanings of the words from context. However, in the ILH studies, participants were provided with the form-meaning connections of words, regardless of whether search was present (e.g., the provision of a paper-based dictionary) or absent (e.g., marginal glosses). Therefore, participants had access to information to learn form-meaning connections in either condition. Studies that have found significant effects of dictionary use and frequency of look-up may provide evidence for the benefit of having access to form-meaning mappings of unknown words (e.g., Ko, 1995), rather than demonstrate the benefit of the cognitive process of search. Nation and Webb's (2017) Technique Feature Analysis, for example, considers whether an activity ensures successful linking of form and meaning as one of the key components for vocabulary learning. Considering whether successful form and meaning links are made by participants may provide a better estimation of learning gains compared to whether learners are provided with the information about words or have to search for it.

Laufer (1999) provided a different coding of search, i.e., conceptualizing search with three levels: (i) no, (ii) moderate—searching for the meaning of a word, and (iii) strong search—searching for the form of a word). However, among the studies that examined tasks with a search component, only one study (Snoder, 2017) included strong search and the rest included moderate search. The effects of strong search compared to moderate search has rarely been investigated, indicating a need for further research on the role of search intensity.

---

[9] It should be noted that Hill and Laufer (2002) and Peters (2007) did not intend to test ILH and search was not operationalized in terms of the search component of the ILH.

2.5.2 Influence of Empirically Motivated Variables on the Effects of Involvement Load

Time on task. The results showed that although time on task was positively correlated with learning gains, this trend disappeared when IL was controlled. Additional analysis found a positive correlation between IL and time on task indicating that the effect of tasks taking longer is mainly due to a greater IL (Hulstijn & Laufer, 2001). This suggests that engaging in a longer task does not necessarily lead to greater learning gains and that the IL of the task would better explain vocabulary learning. This might best be illustrated by comparing sentence writing and reading with glosses. Reading with glosses can take longer because of the length of the text, but learners may spend their time focused on understanding the text rather than paying attention to target words. In contrast, in sentence writing students have greater involvement with each target word, which in turn contributes to greater learning gains. In this case, although reading with glosses (IL = 1: moderate need) takes longer than writing (IL =3: moderate need and strong evaluation), it is less effective as indicated by its IL.

Frequency. The results indicated that frequency positively contributed to learning on immediate posttests with the estimated learning gain increasing by 8.3% as frequency increased by 1 when controlling the effect of IL. This highlights the importance of frequency on vocabulary learning as well as the quality of processing (Schmitt, 2008; Webb & Nation, 2017). One explanation of the frequency effect is that it provides retrieval opportunities for students (Folse, 2006). When students encounter a word repeatedly while reading, they are likely to focus greater attention on unfamiliar words in the first several encounters to try to infer their meanings and retrieve information learned about that word from the previous encounters (Uchihara et al., 2019; Rott, 2007; Webb, 2007). Similarly, when students engage in fill-in-the-blanks activities where they must use target words repeatedly, they may try to retrieve the forms of the words that they used previously (Webb & Nation, 2017). Since the ILH only focuses on the process of learning unknown words, it does not consider retrieval opportunity as a component (Laufer, 2020). However, this finding suggests that including retrieval opportunity may enhance the prediction of task effectiveness (Nation & Webb, 2011).

The frequency effect was not found on delayed posttests. This may be due to limited frequencies in the included studies. Studies exploring the effect of frequency on incidental vocabulary learning from reading have shown that many encounters are required for sizable learning to occur (e.g., Elgort & Warren, 2014; Pellicer-Sánchez & Schmitt, 2010; Waring & Takaki, 2003). Among the studies included in this meta-analysis, the mean frequency was 1.5 (*SD* = 1.16, Median = 1, Min = 1, Max =8) and the majority of the ESs (81.5%) were from the learning conditions where target words were not repeated. When tasks included repetition, the mean frequency was also quite low (3.69, *SD* = 1.17, Median = 4). To have a meaningful impact on retention, a higher number of repetitions may be required.

No interaction between frequency and IL was found. This suggests that the effect of IL may not change in relation to frequency and contrasts with earlier suggestions that the effect of IL decreases as frequency increases or that frequency is more important than the IL of a task (Folse, 2006; Hulstijn & Laufer, 2001). Instead, this finding indicates that frequency influences learning independently of IL. However, this result must be interpreted with caution. Only four studies (Ansarin & Bayazidi, 2016; Folse, 2006; Jahangiri & Abilipour, 2014; Y.-T. Lee & Hirsh, 2012) explicitly investigated the interaction between frequency and IL, and all of these studies included only two different frequencies (e.g., 1 time vs. 3 times). This points to the possibility that this meta-analysis did not have enough data to accurately assess the interaction effect. To draw a clearer conclusion on whether the influence of IL changes as the frequency of encounters/use increases, more studies directly examining the interaction between frequency and the ILH are needed.

Aspect of vocabulary knowledge. The results showed that aspects of vocabulary knowledge develop differently through incidental vocabulary learning tasks. The overall learning gain of the ILH studies was the highest for form knowledge, followed by form-meaning recognition, form-meaning recall, use, and VKS, in that order. Following previous findings, form knowledge develops first, followed by form-meaning connections (e.g., Webb, 2007, see also, Schmitt, 2000). Knowledge of use may be more difficult to gain since learners need to acquire different types of lexical information (e.g.,

collocational knowledge, grammatical knowledge in addition to form-meaning connection) and limited processing resources may restrict learning to the aspects of knowledge that receive attention (Barcroft, 2015).

The analyses of the interaction between IL and vocabulary knowledge yielded different results between immediate and delayed posttests. On immediate posttests, the influence of IL was the most pronounced for (1) form-meaning recall, followed by (2) use, (3) form-meaning recognition, (4) form, and (5) VKS, in that order. On delayed posttests, the influence of IL was strongest for (1) form, followed by (2) form-meaning recognition, (3) form-meaning recall, (4) use, and (5) VKS in that order. This difference might suggest that on immediate posttests, relatively demanding knowledge such as form-meaning recall and use distinguishes the effect of IL more clearly, while on delayed posttests, more easily gained aspects of vocabulary knowledge such as form and form-meaning recognition can capture the influence of IL better. This could be explained by the interaction between the sensitivity of test formats and learning decay over time. When learning gains are measured immediately after learning with more sensitive tests (i.e., recognition tests), learners may easily recognize target words they were exposed to even during a low IL task. Therefore, the differences in learning gains across tasks may be less pronounced compared to the delayed posttest, where learning gains tend to show decay over time and differences in gains across activities may be revealed by more sensitive tests. Form-recognition and form-meaning recognition test formats are sensitive to smaller degrees in knowledge and may have captured differences in gains that less sensitive tests such as form-meaning recall and use tests cannot capture.

Both on immediate and delayed posttests, VKS showed the least sensitivity to IL. One potential explanation is that VKS is not sensitive enough to capture the influence of task features on learning because VKS lumps different aspects of vocabulary knowledge together to calculate a single score. VKS has been criticized for its ambiguity in what the test score represents (e.g., Schmitt, 2010) and this characteristic might have blurred the effect of IL. However, given the fact that (1) no statistical significance was found between all combinations and (2) relatively small numbers of ESs for studies measuring

use, form-meaning receptive recognition, and form knowledge, these findings should be interpreted with caution.

L2 proficiency. The results indicated that participants' L2 proficiency may not influence learning or the effect of IL. This suggests that (1) learning gains might not have differed based on learners' L2 proficiency and (2) learners might have benefitted from IL to similar degrees regardless of their proficiency. It may be that once less proficient learners reach a proficiency with which they can complete language tasks adequately, they can benefit from tasks with higher ILs in a manner similar to more advanced learners (Kim, 2008). However, it is also important to consider that researchers and instructors would most likely have used tasks and target words that they deemed to be appropriate for the participants' level of proficiency. Thus, the effects of proficiency on vocabulary learning may not be reflected in the sample of studies examined.

2.5.3 Suggestions for Future Research

Future individual studies. This meta-analysis identified several factors that need investigation to deepen our understanding of how the components of the ILH and other factors contribute to vocabulary learning. First, none of the studies included a learning condition with strong need where learners select certain unknown words to pursue the goals of their tasks. Although motivational factors on vocabulary learning have occasionally been discussed (e.g., Tseng & Schmitt, 2008), few studies have examined how different manipulations of motivational factors influence the effectiveness of tasks. Future research needs to look further at how motivational factors affect learning by examining learning conditions with varying degrees of the need component.

Second, most studies focused on single word learning (However, see Cao, 2013; Snoder, 2017), making it difficult to draw a conclusion on the predictive ability of the ILH in terms of multiword item learning. Similarly, most studies had either no repetitions or a small number of repetitions making it difficult to clarify how frequency interacts with the effect of conditions present in tasks. Additionally, more studies need to investigate the relationship between vocabulary knowledge and the ILH by measuring different aspects of vocabulary knowledge. Learning was mainly measured using form-

meaning connection tests (e.g., translation tests or multiple-choice tests). Although the use of other test formats was observed, it is still not clear how other components of vocabulary knowledge such as collocations, associations, spelling, pronunciation, and constraints on use (e.g., Webb, 2005) develop through engaging in tasks.

Future meta-analyses. First, it would be useful to compare the ILH to other frameworks that make predictions about L2 vocabulary learning. For example, Nation and Webb's (2011) Technique Feature Analysis (TFA) considers other factors that are reported to contribute to vocabulary learning such as retrieval, interference, and negotiation. Furthermore, there are factors reported to contribute to learning not included in the ILH or TFA such as use of chunking, hierarchical organization, pre-task planning (Zou, 2017) and mode of input (e.g., Feng & Webb, 2019; Vidal, 2011). Comparing the ILH and TFA, as well as examining other reported factors may reveal a more comprehensive picture of how learning conditions contribute to vocabulary development and enable a more accurate prediction of task effectiveness.

Second, this meta-analysis exclusively focused on the predictive ability of the ILH within the realm of incidental vocabulary learning. However, it may also be useful to examine how the ILH predicts learning in deliberate vocabulary learning activities (e.g., the keyword technique, flashcard learning, and crossword puzzles). For example, although this study did not find that search contributes to incidental learning, search might positively affect deliberate learning. This is because when information about target words is not at learners' disposal, learners may try to retrieve it from their memory and this retrieval attempt potentially enhances learning (Nation & Webb, 2011). Such an application of the ILH had been discussed previously (Nation & Webb, 2011), but was never systematically analyzed by looking at learning gains reported in earlier studies. Hence, it may be useful for future meta-analyses to look at the predictive power of the ILH on vocabulary learning in wider contexts. Additionally, Laufer and Hulstijn (2001) claim that in intentional learning, the IL effects and individual learning strategies are confounded. Therefore, if the ILH is to be studied for intentional learning, it may be important to ensure that IL and individual strategies are disentangled. One way to do so is

to conduct within-subject studies, where the effects of different intentional learning tasks can be compared while controlling for the strategies used by each participant.

Lastly, in addition to the ILH and TFA, there are several conceptualizations of factors that contribute to vocabulary learning. Schmitt's (2008) discussion of *engagement* might be a complementary concept that may help explain learning. He argued "[i]n essence, anything that leads to more and better engagement should improve vocabulary learning, and thus promoting engagement is the most fundamental task for teachers and materials writers, and indeed, learners themselves (Schmitt, 2008, p. 339–340)". He listed nine factors facilitating vocabulary learning: (1) increased frequency of exposure; (2) increased attention focused on the lexical item; (3) increased noticing of the lexical item; (4) increased intention to learn the lexical item; (5) a requirement to learn the lexical item (by teacher, test, syllabus); (6) a need to learn/use the lexical item (for task or for a personal goal); (7) increased manipulation of the lexical item and its properties; (8) increased amount of time spent engaging with the lexical item; (9) amount of interaction spent on the lexical item. More recently Webb and Nation (2017) described quality of attention as including four factors (noticing, retrieval, varied encounters and use, and elaboration) to help explain how learning may occur within and across activities. To improve our understanding of the conditions that contribute to vocabulary learning, it may be useful for future studies to code for empirically motivated factors that are present in their learning conditions. This would provide a much larger amount of more transparent data that could then be examined simultaneously in future meta-analyses.

2.6 Conclusion

The most important contribution of the ILH to vocabulary research might be that it builds a model of vocabulary learning by focusing on multiple factors simultaneously. Before the ILH, individual studies tended to only focus on one or a limited number of factors influencing vocabulary learning. Although this focused approach is important, researchers' discussions were often restricted to whether a given factor (or certain learning conditions) was useful for learning or not. The ILH enabled researchers to compare multiple factors across different learning conditions by providing a falsifiable

hypothesis. This provided not only a theoretical contribution to the field but also enabled more accurate and practical pedagogical suggestions by investigating the relative values of different factors.

The current meta-analysis synthesized the results of the studies that strictly controlled IL. The findings supported ILH's prediction by revealing a clear trend showing that higher ILs led to greater learning gains. Additionally, the findings suggested some potential for the ILH to be improved. A large variance in learning gains remained unexplained by the ILH. Different components had varying contributions to learning, i.e., evaluation had the greatest influence, followed by need, while search had little effect. With these findings, the current study makes a step toward enhancing the hypothesis to better explain vocabulary learning and provide more empirically based pedagogical suggestions.

There have been several discussions of theories of L2 vocabulary learning (e.g., Barcroft, 2015; Dóczi & Kormos, 2016; Hulstijn, 2001; Kormos, 2020; Laufer, 2020; Moonen et al., 2006; Nation, 2013; Nation & Webb, 2011; see also Suzuki et al., 2020 for a recent discussion of desirable difficulty; Lightbown, 2008, for Transfer Appropriate Processing). However, (quasi-) empirical studies aiming to directly contribute to theory building are relatively scarce with the majority of these studies focusing on the ILH (but see also Barcroft, 2002, 2003, 2004, 2009, 2019; Kida & Barcroft, 2018, testing the type of processing – resource allocation (TOPRA) model). The large number of studies investigating the ILH may be due to its strengths: (i) proposing a clear falsifiable hypothesis, (ii) demonstrating how the hypothesis can be tested (Hulstijn and Laufer, 2001), and (iii) aiming to provide transparent pedagogical suggestions—the findings of the ILH studies provide pedagogical implications that can easily be applied to vocabulary teaching. In order to further develop models explaining vocabulary learning, it would be useful to carry out more studies (i) directly contributing to theory building by testing hypotheses (i.e., the ILH and other related hypotheses of vocabulary learning), (ii) comparing different hypotheses, and (iii) synthesizing those findings comprehensibly.

2.7 References

The full reference list of the studies included in the meta-analysis is available in Appendix G.

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*(2), 135–146. https://doi.org/10.2989/16073614.2016.1201774

Baddeley, A. D. (1978). The trouble with levels: A reexamination of Craik and Lockhart's framework for memory research. *Psychological Review*, *85*(3), 139–152. http://dx.doi.org/10.1037/0033-295X.85.3.139

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. 1–22). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58542-0_42-1

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, *52*(2), 323–363. https://doi.org/10.1111/0023-8333.00186

Barcroft, J. (2003). Effects of questions about word meaning during L2 Spanish lexical learning. *The Modern Language Journal*, *87*(4), 546–561. https://doi.org/10.1111/0023-8333.00186

Barcroft, J. (2004). Effects of sentence writing in second language lexical acquisition. *Second Language Research*, *20*(4), 303–334. https://doi.org/10.1191/0267658304sr233oa

Barcroft, J. (2009). Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Quarterly*, *43*(1), 79–103. https://doi.org/10.1002/j.1545-7249.2009.tb00228.x

Barcroft, J. (2015). *Lexical Input Processing and Vocabulary Learning*. Amsterdam: John Benjamins

Barcroft, J. (2019). Sentence-level processing for content and new L2 words: Where does deeper processing go? In R. P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 242–257). New York, NY: Routledge.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348–393. https://doi.org/10.1111/lang.12224

Cao, Z. (2013). The effects of tasks on the learning of lexical bundles by Chinese EFL learners. *Theory and Practice in Language Studies*, *3*(6), 957–962. https://doi.org/10.4304/tpls.3.6.957-962

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: Guilford Press.

Cheng, H.-C. (2011). *Vocabulary acquisition in learning English as a second language: Examining the involvement load hypothesis and language anxiety with Taiwanese college students* (Unpublished doctoral dissertation, University of Northern Colorado). University of Northern Colorado, Greeley, Colorado.

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

Cho, K.-S., & Krashen, S. D. (1994). Acquisition of vocabulary from the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading*, *37*(8), 662–667.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. http://dx.doi.org/10.1037/0096-3445.104.3.268

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

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford: Oxford University Press.

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*(2), 227–252. https://doi.org/10.1177/1362168811431377

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*(2), 365–414. https://doi.org/10.1111/lang.12052

Feng, Y., & Webb, S. (2019). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, 1–25. https://doi.org/10.1017/S0272263119000494

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01373-9

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second language learning*. Rowley, MA: Newbury House.

Hazrat, M. (2015). The effects of task type and task involvement load on vocabulary learning. *Waikato Journal of Education*, *20*(2), 79–92. https://doi.org/10.15663/wje.v20i2.189

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press. https://doi.org/10.1016/B978-0-08-057065-5.50001-4

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

Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *International Review of Applied Linguistics in Language Teaching; Heidelberg*, *41*(2), 87–106. https://doi.org/10.1515/iral.2003.007

Hirata, Y., & Mori, C. (2008). A study of effective tasks based on task-induced involvement in incidental vocabulary acquisition. *International Journal of Curriculum Development and Practice*, *10*(1), 25–37. https://doi.org/10.18993/jcrdaen.10.1_25

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

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*(4), 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: Cambridge University Press.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Hyun, P. J. (2011). *The role of task-induced involvement load in vocabulary acquisition of Korean college students* (Unpublished master's thesis, Ewha Womans University). Ewha Womans University.

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.

Jahangiri, K., & Abilipour, I. (2014). Effects of collaboration and exercise type on incidental vocabulary learning: Evidence against involvement load hypothesis. *Procedia - Social and Behavioral Sciences*, *98*, 704–712. https://doi.org/10.1016/j.sbspro.2014.03.471

Kaivanpanah, S., & Miri, M. (2018). Inspecting task-induced involvement from the perspective of sociocultural theory. *Journal of Teaching Language Skills*, *37*(1), 159–192. https://doi.org/10.22099/jtls.2019.30652.2569

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, *12*(3), 365–386. https://doi.org/10.1177/1362168808089922

Keyvanfar, A., & Badraghi, A. H. (2011). Revisiting task-induced involvement load and vocabulary enhancement: Insights from the EFL setting of Iran. *Man & the Word/Žmogus Ir Žodis*, *13*(3), 56–66.

Kida, S., & Barcroft, J. (2018). Semantic and structural tasks for the mapping component of L2 vocabulary learning. *Studies in Second Language Acquisition*, *40*(03), 477–502. https://doi.org/10.1017/S0272263117000146

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, *58*(2), 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal*, *78*(3), 285. https://doi.org/10.2307/330108

Ko, H. M. (1995). Glossing in incidental and intentional learning of foreign language vocabulary and reading. *University of Hawai'i Working Papers in ESL*, *13*(2), 49–94.

Konno, K., Takanami, S., Okuyama, Y., & Hirai, A. (2009). Examining the effects of involvement load on Japanese EFL learners' vocabulary retention. *JLTA Journal*, *12*, 46–64. https://doi.org/10.20622/jltaj.12.0_46

Kormos, J. (2020). How does vocabulary fit into theories of second language learning? In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 207–222). Routledge. https://doi.org/10.4324/9780429291586-14

Laufer, B. (1999). Task effect on instructed vocabulary learning: The hypothesis of "involvement." *Selected Papers from AILA'99 Tokyo*, 47–62. Tokyo: Waseda Univiversity Press.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. https://doi.org/10.1191/0265532204lt277oa

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

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

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). Clevedon, UK: Multilingual Matters.

Maftoon, P., & Haratmeh, M. S. (2012). The relative effectiveness of input and output-oriented tasks with different involvement loads on the receptive and productive vocabulary knowledge of Iranian EFL learners. *The Journal of Teaching Language Skills*, *4*(2), 27–52.

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

Moonen, M. L. I., De Graaff, R., & Westhoff, G. J. (2006). Focused tasks, mental actions and second language teaching: Cognitive and connectionist accounts of task effectiveness. *ITL - International Journal of Applied Linguistics*, *152*(0), 35–55. https://doi.org/10.2143/ITL.152.0.2017862

Nakata, T., & Webb, S. (2017). Vocabulary learning exercises: Evaluating a selection of exercises commonly featured in language learning materials. In B. Tomlinson, University of Liverpool, & Materials Development Association (United Kingdom) (Eds.), *SLA research and materials development for language learning*. New York: Routledge.

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

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

Nation, P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 15–29). Routledge. https://doi.org/10.4324/9780429291586-2

Newton, J. (2020). Approaches to learning vocabulary inside the classroom. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 255–270). Routledge. https://doi.org/10.4324/9780429291586-17

Nguyen, C.-D., & Boers, F. (2018). The effect of content retelling on vocabulary uptake from a TED talk. *TESOL Quarterly*. https://doi.org/10.1002/tesq.441

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning*, *50*(1), 57–85. https://doi.org/10.1111/0023-8333.00111

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., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language*, *22*(1), 31–55.

Peters, E. (2007). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning & Technology*, *11*(2), 36–58.

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques

compared. *Language Learning*, *59*(1), 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (1st ed., pp. 23–45; By L. Plonsky). Routledge. https://doi.org/10.4324/9781315870908-3

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

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

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

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, *57*(2), 165–199. https://doi.org/10.1111/j.1467-9922.2007.00406.x

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). New York: Cambridge University Press.

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

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

Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of involvement load, spacing, and intentionality. *TESL Canada Journal, 34*(3), 140–164. https://doi.org/10.18806/tesl.v34i3.1277

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, *103*(3), 713–720. https://doi.org/10.1111/modl.12585

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

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.

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

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

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*(2), 357–400. https://doi.org/10.1111/j.1467-9922.2008.00444.x

Tsubaki, M. (2012). *Vocabulary learning with graphic organizers in the EFL environment: Inquiry into the involvement load hypothesis* (Unpublished doctoral dissertation). Temple University, Philadelphia, PA.

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

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 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. *J Stat Softw*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

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*(3), 59–66. https://doi.org/10.5539/ijel.v4n3p59

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 1–17.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52. https://doi.org/10.1017/S0272263105050023

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

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13–40. https://doi.org/10.3138/cmlr.53.1.13

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

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*(1), 54–75. https://doi.org/10.1177/1362168816652418

Chapter 3

3 Updating the Involvement Load Hypothesis: Creating an improved predictive model of incidental vocabulary learning

3.1 Introduction

Laufer and Hulstijn's (2001) Involvement Load Hypothesis (ILH) was designed to predict the effectiveness of instructional activities 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, 2019; 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 ILH predictions 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, test format) should also be included (e.g., Folse, 2006). To evaluate the predictive ability of the ILH, Author (XXXX) adopted a meta-analytic approach to statistically summarize studies that tested 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. The findings raised the possibility that the predictive ability of the ILH could be enhanced by evaluating the relative influence of each ILH component and including other empirically motivated factors that affect incidental vocabulary learning. Therefore, the present study aims to determine whether it is possible to improve the ILH to enhance its accuracy in predicting incidental vocabulary learning.

3.2 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 involving higher degrees of these factors are predicted to have 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: (i) absent when the unknown word is not needed to complete the activity (0 points), (ii) moderate when an external entity (e.g., activity or teacher) asks students to understand or use the word (1 point), and (iii) strong when the need for the word is derived by the learners themselves, e.g., 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 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 activities that include 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 the word to use (0 points). It is moderate when a 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 sentence production activities (2 points).

The involvement load (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 a clear intention to commit 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 target words and learning gains may be reflected by the strategy that was used instead of the learning activity they engaged in.

The ILH claims that when *other factors are equal*, words which are processed with higher involvement load will be retained better than words which are processed with 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 is also be useful to consider whether the addition of other factors might enhance the accuracy of the ILH. 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 to a wider variety of contexts.

3.2.1 Earlier Studies Testing the ILH Predictions

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) or 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 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 ILH prediction. 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 ILH prediction.

In contrast, many studies yielded findings that were not entirely in line with the ILH 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. Keating (2008) found that while the results on an immediate posttest supported ILH, the results on the same test 2 weeks later only provide partial support. 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, Koizumi, Tomita, 2020). Author (XXXX) meta-analyzed earlier studies that tested the ILH prediction. 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 referring only to the IL of an activity has limited accuracy in predicting learning gains. The meta-analysis also revealed that the individual components of the ILH (need, search, evaluation) contributed to learning at varying degrees. Evaluation was found to contribute to the greatest amount of learning, followed by need. However, search was not found to contribute to learning. These findings challenge the assumption of the ILH that each component influences learning to the same degree and raises the possibility of enhancing its prediction by investigating their degrees of influence.

3.2.2 Potential Approaches to Enhancing the ILH

Results of earlier studies testing the prediction of the ILH suggest potential approaches to enhancing the accuracy of ILH predictions. These approaches may include: (a) evaluating the degree of influence of each ILH component, (b) revising the evaluation component, and (c) adding other factors to the ILH.

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 the different 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. Laufer and Hulstijn (2001) mentioned this possibility and recommended further investigation of the influence of each component. Several studies have also indicated that the 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. Author’s (XXXX) meta-analysis of the ILH captured this trend revealing that evaluation had the most substantial influence, followed by need, while search was found to have no influence on

learning. It is also important to note that the ILH assumes that strong need and strong evaluation have double the impact on learning as do moderate need and evaluation (2 points are awarded for both of strong need and evaluation, while 1 point is awarded for moderate need and evaluation). It remains to be determined whether this is indeed the case. Therefore, it would be useful to separately examine the degree of influence of each level of the individual components to enhance the accuracy of ILH predictions.

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 ILH prediction while manipulating the frequency of encounters or use of target words (e.g., Eckerth & Tavakoli, 2012; Folse, 2006; Y.-T. Lee & Hirsh, 2012). Folse (2006) found that an activity with lower IL but repetition of target items contributed to greater vocabulary learning than an activity with 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 ILH prediction 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 ILH 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.

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, 2001), suggest that processing information in visual and verbal channels leads to better retention of target items than processing in either channel alone. Given that in activities such as Jahangard, (2013), Hazrat (2015), and Karalik and Merç (2016) that incorporate speaking and listening, students are often provided with the written and spoken forms of target words, Multimedia Learning Theory and Dual Coding Theory would suggest that these activities would 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, (i) recognition (multiple-choice questions) and (ii) recall (meaning and form cued recall tests). Yanagisawa, Webb, and Uchihara (2020) added an *other test format* 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.

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) as a factor might enhance the accuracy of the ILH 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, Boers, & Housen, 2013; Pellicer-Sánchez, 2016). There is 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 ILH predictions.

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 in order 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 enhance 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.

3.2.3 The Current Study

Research has indicated that it would be useful to try to improve upon Laufer and Hulstijn's (2001) ILH framework. Authors' (XXXX) 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 (Author, XXXX; 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). A third way to improve the ILH might be to include other empirically motivated factors (e.g., frequency, mode, test format, test day) to further enhance the accuracy of the prediction (Folse, 2006; Hazrat, 2015; Rott, 2012).

The present study aims to determine whether it is possible to improve the ILH to enhance its accuracy in predicting incidental vocabulary learning. Through meta-analyzing studies examining incidental vocabulary learning gains while strictly controlling the ILs of tasks, we seek to identify the optimal statistical model that best

predicts learning gains. The resulting model will serve as an updated ILH. The updated ILH may be useful for language educators and material writers when choosing and designing effective activities for their students.

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?

3.3 Method

3.3.1 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 the 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). Lastly, the reported learning gains were analyzed using a three-level meta-regression model (Cheung, 2014). The analysis procedure includes (1) identifying the best operationalization of the ILH, **(**2) identifying the best grouping of test formats, and (3) determining the optimal combination of variables that best predicts learning gains. Additionally, we carried out sensitivity analyses to evaluate the robustness of our results.

3.3.2 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, book chapters were also included (Oswald and Plonsky, 2010). Research reports published from 2001 to April 2019 were searched 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, 2012). 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 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 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 A for basic information about the studies).

3.3.3 Dependent Variable: Effect Size Calculation

In order 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, Cobb, & Meara, 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 R statistical environment (R Core Team, 2017). Each calculated ES was weighted using the sampling variance of the posttests scores (see Appendix D for the detailed calculation formulas for ES and sampling variance).

3.3.4 Predictor Variables

We coded the studies for predictor variables: 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 C and H for the details of 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, and (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-, letter-writing). To more clearly distinguish between the different levels of evaluation, 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).

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), (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 & form recall) vs. recognition (meaning recognition & form recognition) vs. other (VKS & use of target words), (c) receptive (receptive recognition & recall) vs. productive (form recognition & form recall) vs. other (VKS & use of target words), and (d) receptive recall vs. productive recall vs. recognition vs. other (VKS & 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 (i) written when participants engaged in a written activity (i.e., reading and writing) or (ii) spoken when participants engaged in a spoken activity (i.e., listening and speaking; e.g., Jahangard, 2013; Hazrat, 2015). 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 inter-coder reliabilities were calculated using Cohen's Kappa coefficient (κ) and the agreement rate was high and acceptable at κ = .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.

3.3.5 Data Analysis

We used a three-level meta-regression model (Cheung, 2014; Lee et al., 2018) to analyze ESs that indicate the proportion of unknown words learned (de Glopper & Swanborn, 1999; Yanagisawa, Webb, & Uchihara, 2020). Three-level meta-regression models can account for different sources of variance (i.e., within- and between-study variances and sampling variance), 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), which potentially increases a Type I error rate. To deal with this, 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 (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 for small sample sizes (AICc, 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: (1) the original ILH model that only includes IL as a single numerical predictor variable (the sum of the scores of the three ILH components), (2) the ILH component model that includes categorical variables denoting each of components (need, search, evaluation) separately for each level (absent, moderate, and strong), and (3) 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 using the identified best ILH operationalization, we fitted four models based on the different groupings of test formats: (i) each test format, (ii) receptive, productive, and other, (iii) recall, recognition, and other, and (iv) receptive recall, productive recall, recognition, and other.

Lastly, 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 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. Since 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 will not be interpreted. Because explained variance is non-negative 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 Appendix I).

3.4 Results

To identify the best combination of predictive variables for incidental vocabulary learning, we first compared different operationalizations of the ILH to determine ILH operationalization that best predicts learning gains. Three-level meta-regression models were fitted with three different operationalizations of the ILH: (1) an original ILH model that only included IL as a single numerical predictor variable (the sum of the scores of the three ILH components), (2) 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 (3) 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 (-150.11 on the immediate posttest and -170.04 on the delayed posttests) followed by the ILH component (-148.04 and -166.21) and the original ILH (-140.02 and -159.00) in that order (Table 1). The calculated Akaike weights also indicated strong support for the modified ILH component model, indicating the probability that this model is the best predictive model among candidate models was 73% on the immediate and 87% on delayed posttests.

Table 1: Comparison of the Different ILH Operationalizations

	Immediate Posttest			Delayed Posttest		
Framework	AICc	ΔAICc	Akaike Weight	AICc	ΔAICc	Akaike Weight
Original ILH model	-140.02	0	0.00	-159.00	0	0.00
ILH component model	-148.04	-8.02	0.26	-166.21	-7.21	0.13
Modified ILH component model	-150.11	-10.09	0.73	-170.04	-11.04	0.87

Note. The smaller the AICc value the better the model. Akaike weight indicates the probability that each model is the best model.

Next, three-level meta-regression models comparing 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 (-198.48 on the immediate and -229.73 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 (-194.96, -226.13), (c) recall vs. recognition vs. other (-187.15, -214.87), and (d) receptive vs. productive vs. other (-165.58, -191.71), in that order (Table 2). This was also strongly supported by the calculated Akaike weights, which indicated the probability was 85% on the immediate and 86% on delayed posttests that the model grouping test format as receptive recall, productive recall, recognition, and others was the best predictive model among all candidate models.

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	-165.58	0	0.00	-191.71	0.00	0.00
Recall vs. Recognition vs. Other	-187.15	-21.56	0.00	-214.87	-23.16	0.00
Each test format	-194.96	-29.38	0.15	-226.13	-34.42	0.14
Receptive Recall vs. Productive Recall vs. Recognition vs. Other	-198.48	-32.90	0.85	-229.73	-38.02	0.86

Note. The smaller the AICc value the better the model. Akaike weight indicates the probability that each model is the best model.

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, spoken) and written was set as 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 six predictors: need, evaluation, sentence-level varied 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 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 21.3% for the inclusion of need ($b = 0.213$, $p = .028$), 8.4% for evaluation ($b = 0.084$, $p = .001$), 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 12.5% and 9.9%, respectively. In contrast, when learning was measured with recognition tests, gains increased by 22.7%. The analyses also showed that learning gains increased by 9.5% as frequency increased by 1 and decreased by 9.8% when mode was spoken (as opposed to written).

All of the included predictors’ influence were confirmed with 95% CIs and *p*-values calculated based on the RVE, except for mode ($p = .093$). Given that model selection referring to AICc 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 examinations to confirm whether it is statistically significant or not (Burnham & Anderson, 2002; see also, Aho et al., 2014).

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	0.070	-0.093	0.232	.377
Test: Productive recall	-0.125	-0.221	-0.030	.022
Test: Recognition	0.227	0.028	0.426	.040
Test: Other	-0.099	-0.158	-0.040	.009
Need	0.213	0.032	0.393	.028
Evaluation	0.084	0.041	0.128	.001
Varied Use (Sentence)	0.153	0.080	0.225	< .001
Varied Use (Composition)	0.233	0.129	0.337	< .001
Frequency	0.095	0.016	0.175	.028
Mode: Spoken	-0.098	-0.226	0.030	.093
Total explained variance	.171			

Between-study variance explained	.000
Within-study variance explained	.594

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.

The resulting model on delayed posttests included seven predictors: need, search, evaluation, sentence-level varied use, and 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 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 14.0% for the inclusion of need (b = 0.140, p = .022), 9% for evaluation (b = 0.090, p = .001), 11.3% for sentence-level varied use (b = 0.113, p < .001), and 20.8% for composition-level varied use (b = 0.208, 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 12.0% and 8.7%, respectively. Whereas, when learning was measured with recognition tests, learning increased by 21.6%. The analyses also showed that learning decreased by 4.9% when search was present. Learning also decreased by 0.4% as the number of days between learning and testing increased by 1.

All of the included predictors were positively related to learning gains, except for test day and search. 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 = .015). The results also showed that when search was present, the estimated mean learning gain decreased by 4.9% (b = 0.49, 95% CI [-0.120, 0.021]). Additionally, p-value calculated based on the RVE showed that search did not reach statistical significance (p = .149),

suggesting that there is great variance in the negative influence of search and it might not necessarily hinder learning, but is 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.

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	0.187	0.063	0.310	.007
Test: Productive recall	-0.120	-0.272	0.031	.092
Test: Recognition	0.216	0.049	0.383	.032
Test: Other	-0.087	-0.128	-0.046	.004
Need	0.140	0.028	0.252	.022
Search	-0.049	-0.120	0.021	.149
Evaluation	0.090	0.043	0.138	.001
Varied Use (Sentence)	0.113	0.060	0.166	< .001
Varied Use (Composition)	0.208	0.155	0.261	< .001
Test day	-0.004	-0.007	-0.001	.015
Total explained variance	.344			

Between-study variance explained	.186
Within-study variance explained	.604

Note. Reading need refers to the need to understand target words while reading. 95% CIs and *p*-values were calculated based on the robust Variance Estimation. For reference level, Test format was set as receptive recall.

The resulting models both on the immediate and delayed posttest also showed greater predictive power than the original ILH as indicated by the increased explained variance at 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 within- and between-study levels explained by the model) (Cheung, 2014). The original ILH model explained 15% of the total variance and 29.1% of the within-study variance on immediate posttests, and 5.1% and 26.5% on delayed posttests. The resulting model explained 17.1% of the total variance and 59.4% of the within-study variance on the immediate posttest, and 34.4% of the total variance and 60.4% of the within-study variance on delayed posttests. 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.

3.5 Discussion

The current study aimed to update ILH through meta-analyzing empirical 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 30.3% (29.1% -> 59.4%) on immediate posttests and by 33.9% (26.5% -> 60.4%) 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 updated ILH provides a more accurate estimation of learning with other factors being equal. Furthermore, the total variance explained also increased by 2.1% (15% -> 17.1%) on the immediate and by 29.3% (5.1% -> 34.4%) on the delayed posttests. This suggests that the updated ILH predicts 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.

3.5.1 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 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 (Laufer & Hulstijn, 2001; Kim, 2008), 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 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 pre-task planning and

hierarchal organization where learners must pay greater attention to the organization of the target words together 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 12.5% on immediate and by 12.0% on delayed posttests) and other test formats (by 9.9% and 8.7%) but increased with recognition (by 22.7% and 21.6%). Given that the type of test greatly influences the 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 the learning gain 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). In immediate posttests, learning gains were estimated to increase by 9.5% as frequency increased by 1 and decrease by 9.8% when mode of input was spoken (as opposed to written). For delayed posttests, learning gains were estimated to decrease by 4.9% when search is present. 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 occurrence or use of target words. Therefore, developing or selecting activities that involve multiple occurrences 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 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, gap-filling). This finding is supported by two earlier studies (Brown et al., 2018; Vidal, 2011) that indicated that reading leads to greater incidental vocabulary learning than listening but contrasted by another that found no difference between the two modes (Feng & Webb, 2019). One reason why reading might contribute to greater learning 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, Webb, & Yanagisawa, 2019; Vidal, 2011). Another explanation could be that L2 learners have a limited capacity for processing L2 spoken input, limited phonological presentations of L2 words (e.g., McArthur, 2003), and smaller oral vocabulary than written vocabulary once their lexical proficiency develops to a certain level (Milton & Hopkins, 2006). In contrast to the results on the immediate posttest, frequency and mode did not contribute to the prediction on the delayed posttest. One plausible explanation is that the positive influences of frequency and mode fade in accord with the retention of learned words as time passes.

The predictive model on the delayed posttests showed test day and search, as well as the ILH components 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 impact of taking immediate posttests on delayed posttests so as to draw a more accurate estimation of the rate that 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 decreased. Authors' (XXXX) 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 literature 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 searching behavior such as dictionary use can be quite demanding for L2 students (Hulstijn, Hollander, & Greidanus, 1996). In contrast, students were provided with information about target words (via glosses or glossaries) in activities without search. Therefore, they may have had more time and opportunities to attend to or process target words by using the forms and meanings of target items provided at their disposal.

Lastly, in contrast to our hypothesis, the number of target words did not clearly contribute to the prediction of learning gains. In order 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 [-0.009, 0.004]), the number of target words was not significantly related to learning gains either on the immediate (p = .205) or the delayed posttests (b = 0.000, 95% CI [-0.007, 0.007], p = .898). 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.

3.5.2 IL Formulas and an Updated ILH

Following the ILH, we created IL formulas based on the resulting models to estimate the relative effectiveness of different incidental learning tasks. Two involvement load formulas were created; one to estimate learning on immediate posttests and the other to estimate retention on delayed posttests.

$$\begin{aligned}
&The\ involvement\ load\ formula\ of\ activities\ for\ immediate\ learning \\
&= [Need\ (absent{:}\ 0\ or\ present{:}\ 1) \times 21.3] \\
&+ [Evaluation\ (0\ or\ 1) \times 8.4] \\
&+ [Sentence\ level\ varied\ use\ (0\ or\ 1) \times 15.3] \\
&+ [Composition\ level\ varied\ use\ (0\ or\ 1) \times 23.3] \\
&+ [Frequency\ (number\ of\ times\ to\ encounter\ or\ use) \times 9.5] \\
&+ [Mode\ (written{:}\ 0\ or\ spoken{:}\ 1) \times -9.8]
\end{aligned}$$

$$\begin{aligned}
&The\ involvement\ load\ formula\ of\ activities\ for\ \ retention \\
&= [Need\ (absent{:}\ 0\ or\ present{:}\ 1) \times 14.0] \\
&+ [Search\ (0\ or\ 1) \times -4.9] \\
&+ [Evaluation\ (0\ or\ 1) \times 9.0] \\
&+ [Sentence\ level\ varied\ use\ (0\ or\ 1) \times 11.3] \\
&+ [Composition\ level\ varied\ use\ (0\ or\ 1) \times 20.8]
\end{aligned}$$

The formulas included seven factors (need, search, evaluation, sentence-level varied use, composition -level varied use, frequency, and mode) to calculate the IL of an activity. The IL of activities express their relative effectiveness for incidental vocabulary learning within studies—when learning gains are measured with the same test format at the same test intervals while dealing with the same participant groups learning the same set of target words. 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 updated ILH:

1. With other factors being equal—when measuring with the same test format at the same timing while dealing with the same set of target words and group of participants, language activities with a higher involvement load will lead to greater incidental word learning than activities with a lower involvement load.
2. Regardless of other factors that are not included in the involvement load formulas, language activities with a higher involvement load will lead to greater incidental word learning than activities with a lower involvement load.

The first hypothesis may be useful for researchers to test the updated ILH as a falsifiable hypothesis in order to evaluate how accurately the updated ILH predicts the relative effectiveness of activities. It may also be useful for educators and material writers to select and design language activities that effectively facilitate vocabulary learning. The second hypothesis is proposed as a null hypothesis. Researchers can examine whether the prediction of the updated ILH holds even when other factors are manipulated while strictly controlling both the factors of the updated ILH and the factor in question. Results from such studies may reveal when the updated ILH does not make accurate estimations and how other factors can be integrated into the updated ILH to further enhance its prediction.

The resulting statistical models' intercept, test format, and test day were not included in the involvement load formulas. This is because although these factors may be useful for calculating the estimated mean learning gains, they are not relevant for making predictions of the order of the effectiveness of activities within studies.

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). Three activities in Laufer and Hulstijn (2001) were (i) reading with glosses, (ii) fill-in-the-blanks, and (iii)

composition writing. All activities included need and frequency as 1. ILs were calculated as 30.8 for reading with glosses, 39.2 for fill-in-the-blanks, and 54.1 for composition writing. When comparing the observed mean test scores in Laufer and Hulstijn (2001), the updated ILH 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 involvement load formula. The ILs were calculated to be 30.8 for reading with glosses, 39.2 for fill-in-the-blanks, 54.1 for composition writing, and 46.1 for sentence writing. Among 24 comparisons of the activities (6 comparisons across 4 activities × 2 test timing × 2 participant groups), the IL formula correctly predicted 22 comparisons (91.7%) of the relative effectiveness between the activities. One thing to note is that when the ILs between activities are similar to each other, the activities are more likely to lead to similar learning gains. For example, the ILs for reading with glosses and fill-in-the-blanks are 30.8 and 39.2, thus learning gains from these activities might be more difficult to detect compared to when comparing learning gains between activities that have greater differences such as composition writing (54.2) and reading with glosses (30.8). Because it is normal for mean scores to fluctuate, there will likely be times when the estimated efficacy order is not observed due to limited statistical power especially when the IL values are close across activities.

Table 5: Coding Examples of the Updated IL (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: $\times \mathbf{21.3}$	1	1	1	1	1	1	1
Evaluation: $\times \mathbf{8.4}$	0	1	0	0	1	0	0
Varied use (sentence): $\times \mathbf{15.3}$	0	0	0	0	0	0	1
Varied use (composition): $\times \mathbf{23.3}$	0	0	1	0	0	1	0
Frequency:	1	1	1	1	1	1	1

$\times \mathbf{9.5}$							
Mode (Spoken): $\times -\mathbf{9.8}$	0	0	0	0	0	0	0
Updated IL	30.8	39.2	54.1	30.8	39.2	54.1	46.1
Order of effectiveness	3	2	1	4	3	1	2

3.5.3 Limitations and Future Directions

First, the updated ILH and IL formulas should be viewed as a simple predictive model. The IL 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 future studies to investigate the predictive accuracy of the updated ILH with learning conditions with 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, 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 theoretic-information 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., 2018; 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 the updated ILH or its involvement load formulas. This is mainly because the limited combinations of factors were investigated in the included studies. However, it might be reasonable to assume 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 example, the effect of frequency might be more pronounced when composition-level varied use was present compared to when evaluation was present (Uchihara et al., 2019). It would be useful for future studies to examine different combinations of factors to examine how these variables interact with each other to influence incidental vocabulary learning.

Lastly, the current study identified some under-researched factors related to the ILH components. None of the meta-analyzed studies included learning conditions with strong need (internal motivation). Search was operationalized only as dictionary use, glossaries, electronic dictionary, and hyperlinked glosses, with no studies examining situations where students guess the meanings of words from context or ask teachers or peers. Future studies should examine these under-investigated conditions to further validate the original ILH and potentially revise the updated ILH model.

3.6 Conclusion

We aimed to update the ILH through meta-analyzing studies examining the ILH. The predictive power of the ILH was improved by (i) examining the influence of each level of individual ILH components, (ii) adopting optimal operationalization of ILH components and test format grouping, and (iii) including other empirically motivated variables. As Box's oft-cited quotation notes "all models are wrong, but some are useful" (Box & Draper, 1987, p. 424), although the updated ILH 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.

Echoing Laufer and Hulstijn (2001), we would like to call for studies to examine the extent to which the updated ILH accurately predicts incidental vocabulary learning gains from language activities. Empirical studies can compare different learning conditions to determine whether the updated ILH 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 the updated ILH is when applied to actual learning contexts. Studies might also investigate other factors that are not included in the updated ILH. Factors might include learner characteristics (e.g., proficiency, Kim, 2008; working memory, Yang, Shintani, & Zhang, 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). Lastly, after accumulating studies that tested the updated ILH, meta-analysis can statistically summarize findings of these studies to revise the updated ILH.

3.7 References

The full reference list of the studies included in the meta-analysis is available in Appendix G.

Author (XXXX) refers to the current thesis's Study 1

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 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(3), 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*(2), 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

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*(2), 136.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag. //www.springer.com/la/book/9780387953649

Chang, A. C.-S. (2019). Effects of narrow reading and listening on L2 vocabulary learning: Multiple dimensions. *Studies in Second Language Acquisition*, 1–26. 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*(2), 211–229. https://doi.org/10.1037/a0032968

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*(2), 227–252. https://doi.org/10.1177/1362168811431377

Feng, Y., & Webb, S. (2019). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, 1–25. https://doi.org/10.1017/S0272263119000494

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 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*(3), 483–517. http://dx.doi.org.proxy1.lib.uwo.ca/10.1017/S0272263113000119

Hazrat, M. (2015). The effects of task type and task involvement load on vocabulary learning. *Waikato Journal of Education*, *20*(2), 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*(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*(2), 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*(4), 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*(3), 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*(3), 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*(1), 169–184. https://doi.org/10.5054/tq.2010.215253

In'nami, Y., Koizumi, R., & Tomita, Y. (2019). Meta-analysis in applied linguistics. In J. McKinley & H. Rose (Eds.), *The Routledge Handbook of Research Methods in Applied Linguistics* (1st ed., 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*(35), 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*(3), 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*(2), 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*(3), 406–428. https://doi.org/10.22158/selt.v5n3p406

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*(4), 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), 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*(1), 89–108. https://doi.org/10.1177/003368829702800106

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

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*(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*(2), 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*(1), 127–147. https://doi.org/10.3138/cmlr.63.1.127

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; New York*, *38*(1), 97–130. http://dx.doi.org.proxy1.lib.uwo.ca/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. https://www.routledge.com/products/9780415718349

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*(4), 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*(3), 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*(4), 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*(3), 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*(1), 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*(3), 261–285. https://doi.org/10.2307/1170540

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. http://link.springer.com/content/pdf/10.1057/9781137449788.pdf#page=140

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

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

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

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 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. *J Stat Softw*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

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*(3), 59–66. https://doi.org/10.5539/ijel.v4n3p59

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

Webb, S., & Chang, A. (2012). Vocabulary learning through assisted and unassisted repeated reading. *Canadian Modern Language Review*, *68*(3), 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*(1), 13–40. https://doi.org/10.3138/cmlr.53.1.13

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

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*(1), 54–75. https://doi.org/10.1177/1362168816652418

Chapter 4

4 What are the predicted learning gains for different incidental vocabulary learning activities?

4.1 Introduction

Research has consistently demonstrated that second language (L2) students can learn vocabulary incidentally (Webb, 2020). Studies have revealed that L2 vocabulary learning occurs through reading (Horst, Cobb, & Meara, 1998; Waring & Takaki, 2003), listening (e.g., Pavia, Webb, & Faez, 2019; van Zeeland & Schmitt, 2013), and viewing (e.g., Rodgers & Webb, 2019). Furthermore, in addition to incidental learning from these meaning-focused input activities, there are many other ways to learn words incidentally. Research examining Laufer and Hulstijn's (2001) Involvement Load Hypothesis (ILH) has also shown that students learn vocabulary as a by-product of completing a variety of language activities such as gap-filling (e.g., Kim, 2008, Folse, 2006), composition writing (Laufer, 2003), sentence writing (e.g., Kim, 2008; Folse, 2006), and retelling (Nguyen and Boers, 2018). These activities involve different cognitive processes that contribute to learning target words incidentally with the presence or absence of these processes affecting learning (Laufer & Hulstijn, 2001).

To predict the relative value of incidental learning activities, Laufer and Hulstijn (2001) created the ILH. The ILH claims that the retention of unknown L2 words is contingent upon the *Involvement Load* (IL) of a task, i.e., the degree to which a task involves one motivational (*need*) and two psychological factors (*search* and *evaluation*). The ILH predicts that activities involving greater IL lead to greater vocabulary learning than activities involving lesser IL. The ILH has been shown to be relatively effective at determining whether one activity is more effective than another activity (e.g., Eckerth & Tavakoli, 2012; Hulstijn & Laufer, 2001; Kim, 2008). Recently, Author (XXXXb) meta-analyzed studies testing the ILH and enhanced its predictive power by fine-tuning the influence of each ILH component (*need, search,* and *evaluation*) and identifying additional variables that contribute to incidental vocabulary learning.

Although studies investigating the ILH have revealed a great deal about the effectiveness of different approaches to word learning, it may be difficult for teachers and learners to apply findings to pedagogy. Research investigating the ILH typically involves the comparison of slightly different learning conditions with studies often revealing slightly different conclusions even when focusing on the same activities (e.g., Folse, 2006; Hulstijn & Laufer; Keating, 2008). Thus, research findings may not be transparent in a way that teachers and learners can easily apply to language learning. However, because studies testing the ILH examined incidental vocabulary learning from different activities while strictly controlling the ILH components of learning conditions, activity types can be categorized according to their included components. The reported learning gains for the different activity types can then be statistically summarized to obtain estimated overall vocabulary learning gains for each activity type. This may provide more transparent findings that can be easily applied to pedagogy.

The present study uses a meta-analytic approach to (a) provide an overview of the different incidental vocabulary learning conditions that have been examined in studies of the ILH, and (b) obtain the estimated vocabulary learning gains occurring across those activities. One of the key contributions of the present study is to make research investigating vocabulary learning more easily applied to pedagogy. The findings may serve as a guideline for the selection of classroom vocabulary learning activities.

4.2 Background

Research has examined different types of activities to determine how features of learning conditions affect incidental vocabulary learning. These activities can roughly be categorized into two groups. The first group consists of meaning-focused input (MFI) activities aimed at promoting incidental learning through exposure to large amounts of L2 input over the long term. In these activities, students are focused on understanding and enjoying the information that is presented. Examples of MFI activities include extensive reading (e.g., Horst, Cobb, & Meara, 1998; Waring & Takaki, 2006), extensive listening (e.g., Pavia, Webb, & Faez, 2019; van Zeeland & Schmitt, 2013), and extensive viewing (e.g., Rodgers & Webb, 2019). The second group of activities might be described as more

traditional classroom-based activities aiming at promoting learning through exposure to smaller amounts of input over a relatively small amount of time. Examples of this approach include reading a short text, gap-filling (Kim, 2008; Folse, 2006), writing (Kim, 2008; Laufer, 2003; Folse, 2006), and multiple-choice activities (Bao, 2015).

Reading activities that promote incidental vocabulary learning can refer to a variety reading conditions such as (a) reading with L1 glosses, (b) reading with L2 glosses, (c) reading with multiple-choice questions, (d) reading with comprehension questions that require the understanding of target words, (e) reading with comprehension questions that do not require the understanding of target words. With these small differences, it is very difficult for researchers, teachers, and learners to apply the results of learning from one activity to practice.

4.2.1 To What Extent are Words Learned Incidentally Through Different Activities?

Research investigating incidental vocabulary learning has produced a range of learning gains. For example, Hulstijn and Laufer (2001) compared three types of activities: reading comprehension with marginal glosses, fill-in-the-blanks in a text, and writing a composition using target words. Learning gains were measured with a meaning recall test. The results of the immediate posttests with Dutch-speakers learning English showed that on average, 27% of target words were learned from reading, 29% from fill-in-the-blanks, and 49% from writing. The same test with Hebrew-speakers learning English showed that the mean learning gains were 20% of words from reading, 40% from fill-in-the-blanks, and 69% from writing. Kim (2008) compared the learning gains from four activities: reading with marginal glosses, fill-in-the-blanks, composition writing, and sentence writing. She measured learning gains with the vocabulary knowledge scale (VKS: Wesche & Paribakht, 1996). The results of the immediate posttest with lower proficiency ESL students showed that the mean percentage of learning gains were 17.6% from reading with glosses, 21.0% from fill-in-the-blanks, 27.4% from composition writing, and 27.5% from sentence writing. The results of the same test with higher proficiency ESL students showed that the mean percentage learning gains were 27.9%

from reading with glosses, 36.3% from fill-in-the-blanks, 51.2% from composition writing, and 43.2% from sentence writing. Folse (2006) also administered VKS to measure incidental vocabulary learning from fill-in-the-blank once, fill-in-the-blanks three times, and writing sentences using target words. The mean percentage of learning gains on the immediate test were 21.8% for one fill-in-the-blanks, 23.9% for three fill-in-the-blanks, and 47.8% for sentence writing.

Taken together, these studies show that learning gains differed from study to study even when the same types of activities were examined. Additionally, within the same study, learning gains differ greatly based on the group of participants. Because of this complexity in research findings, it is difficult for language teachers and learners to apply findings to language learning. Furthermore, learning conditions with similar labels tend to be slightly different. The effect of writing activities was examined with sentence writing and composition writing activities. Thus, it is not clear whether these activities can be lumped together when considering how much L2 vocabulary students learn from the activities. When we have a lot of slightly different activities with similar labels, it can be very challenging to apply the results of any of these activities to practice. Thus, there is a need to have an objective classification of activities. One way to do this is to classify activities according to learning conditions within the activities that contribute to learning gains. Many researchers have talked about the quality of attention and processing of vocabulary as the key contributor to the learning (e.g., Laufer & Hulstijn, 2001; Webb & Nation, 2017). Classifying activities around the factors contributing to learning may allow us to better apply research findings to practice.

4.2.2 Involvement Load Hypothesis

With the variation in learning gains among activities, it is difficult to understand the relative efficacy of those activities. Aiming to explain the relative effectiveness of different learning conditions, Laufer and Hulstijn (2001) created the ILH. The ILH suggests that when learners pay more attention to unknown words and process words in an elaborated manner, these words are more likely to be recalled later. The ILH claims that the retention of new L2 words is contingent upon an activity's *Involvement Load*

(IL), i.e., the extent to which learning conditions include three components: one motivational component (*need*, the necessity to understand or use a word) and two cognitive components (*search*, to look for information about a word, and *evaluation*, the comparison of the information about word meanings or forms). The ILH predicts that language activities with higher ILs lead to greater vocabulary learning than activities with lower ILs.

Many studies have tested the ILH to examine whether it provides accurate predictions of incidental vocabulary learning (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2008; Folse, 2006; Rott, 2012; Zou, 2017). The majority of studies provided general support for the ILH by finding that activities with higher ILs tend to lead to greater vocabulary learning gains than activities with lower ILs. To provide a more objective and reliable summary of the extent to which ILH accurately predicts incidental vocabulary learning, Author (XXXXa) meta-analyzed 42 studies that tested ILH predictions. The results showed that the ILH was significantly predictive of learning gains, indicating a clear positive correlation between vocabulary learning gains and ILs of activities. The results also revealed that the ILH only explained a limited proportion of the variance in effect sizes (ESs) of learning gains. The ILH explained 15.4% and 5.5% of the total variance on the immediate and delayed posttests, respectively, and 29.1% and 26.5% of the within-study variance—which reflects the difference in ESs within each study—, on the immediate and delayed posttests, respectively.

More recently, Author (XXXXb) adopted a model selection approach (Burnham & Anderson, 2002) to compare several statistical models to identify the set of predictor variables that best predicts incidental vocabulary learning gains. Following previous suggestions for the ILH, Authors (XXXXb) aimed to enhance ILH's predictive ability by (i) considering the relative effects of the ILH components (e.g., Kim, 2008; Laufer & Hulstijn, 2001), (iii) distinguishing different types of strong evaluation—sentence-level varied use (i.e., using each target word in a sentence) and composition-level varied use (i.e., using a set of target words in a composition) (Zou, 2017)—, and (iii) adding other predictor variables (i.e., frequency, mode, test format, test day; e.g., Folse, 2006; Jahangard, 2013; Keating, 2008). The following factors were identified as useful

variables to predict learning gains on immediate posttests: need, evaluation (previously labeled as *moderate evaluation*), sentence-level varied use, composition-level varied use, frequency, mode, and test format. Useful predictor variables on delayed posttests were need, evaluation, sentence-level varied use, composition-level varied use, test format, test day (i.e., the number of days between learning and testing), and search (which negatively influences learning). These resulting models were able to account for greater amounts of variance in ESs. The updated ILH explained 17.1% and 34.3% of the total variance in the immediate and delayed posttests, respectively, and 59.4% and 60.4% of the within-study variance, in the immediate and delayed posttests, respectively.

Both the ILH and updated ILH were proposed as falsifiable hypotheses to predict the relative effects of different learning conditions. Although this has great value, it might be challenging to apply the research findings to teaching practice. Author (XXXXa) found that researchers have had difficulty applying the ILH to the coding of incidental vocabulary learning activities; 11 out of 52 studies coded the ILH components differently from Laufer and Hulstijn's (2001) description of the ILH. Additionally, neither the ILH nor updated ILH predicts the extent to which words can be learned in different activity types. This may be preventing language teachers and learners from benefiting from the accumulated research findings.

One way to determine the proportion of words learned in activities is through the use of a meta-analytic approach. Meta-analysis can statistically summarize the results of empirical studies to obtain the overall mean learning gains of different conditions. As in earlier meta-analyses of incidental vocabulary learning from L1 reading (Swanborn & de Glopper, 1999) and L2 glossed reading (Yanagisawa, Webb, & Uchihara, 2020), the reported posttest scores can be converted into ESs that indicate the proportion of unknown words learned (a.k.a. relative learning gains, e.g., Horst et al. 1998; Webb & Chang, 2015) and examined. Meta-analysis provides a more reliable estimation of how many unknown words can be learned for different groups of learning conditions by looking at the results of multiple studies. Furthermore, because meta-analysis can account for the variance in learning gains, it allows us to calculate predictor intervals for the estimated learning gains. Predictor intervals indicate the range within which future

learning gains will fall with a certain probability. The estimated mean learning gains and their predictor variables for different activity types, each of which includes the same factors that contribute to vocabulary learning, may reveal the extent to which vocabulary learning is likely to occur through engaging in different activities.

4.2.3 Current Study

The current study aims to investigate the degree to which L2 students can incidentally learn new words by engaging in language activities. We adopted a meta-analytic approach to categorize the different activities that have been used to incidentally learn vocabulary and statistically summarize the learning gains that were reported in studies examining the ILH. There are several advantages to meta-analyzing the ILH studies.

First, the learning conditions in studies of the ILH were strictly controlled for their ILH components (need, search, and evaluation) and other variables (e.g., frequency, mode, test formats). Therefore, variables that potentially influence learning can be coded and examined in a relatively transparent manner. Variables that can change regardless of activity types (e.g., frequency, mode, and test formats) can also be included as covariates when analyzing with meta-regression analysis to further enhance the estimation of the relative effectiveness of different activities.

Second, studies investigating ILH compared the effectiveness of many different activities. Therefore, the learning conditions presented in the ILH studies provide a relatively broad representation of incidental vocabulary learning activities. Furthermore, we adopted three-level meta-regression models (Cheung, 2014; Hox, 2010) for this study. This statistical technique can account for differences in reported learning gains within individual studies. While taking advantage of studies comparing different activities, this technique enables more reliable and powerful examinations of the relative effects of different activity types while controlling for potential bias. For example, three-level meta-regression models can control for potential bias related to an unbalanced number of studies examining certain types of activities or the potential influence related to each study (e.g., the characteristics of target words and participants).

A third advantages relates to the classification of activity types. Although activities could simply be grouped according to their labels (e.g., reading, writing), there tends to be a lot of variation among similar activities, as well as the cognitive factors that contribute to learning within those activities. For example, there are a variety of reading activities such as reading with marginal glosses, reading with multiple-choice glosses (i.e., learners have to select a meaning that fits in the context), and reading while answering comprehension questions that require students to understand unknown target words. Grouping these different learning conditions as one activity type (*reading*) might provide a misleading account of vocabulary learning, because the activities include different cognitive processes. For example, reading with marginal glosses provides students with the meanings of target words, whereas reading with multiple-choice glosses requires students to read a text carefully to select the meaning that best fits the context (e.g., Rott, 2005). Similarly, reading while answering comprehension questions related to target words is likely to direct students' attention to target words. Given the different cognitive processes involved in these activities, they might lead to different learning gains. In contrast, activities can be grouped according to the cognitive processing of target words indicated by the updated ILH. For example, reading with and without reading comprehension questions can be distinguished as two different types of activities; the former includes the need component while the latter does not. Grouping learning conditions following their ILH components allows objective categorization of activities based on the cognitive processing of target words.

The following questions guided the current study.

3. What are the different incidental vocabulary learning activities that have been examined in studies of the ILH?
4. What are the estimated mean learning gains that occur through completing different incidental vocabulary learning activities?

4.3 Method

4.3.1 Research Design

To obtain the mean learning gains for different incidental vocabulary learning activities, we adopted a meta-analytic approach to statistically summarize the results of earlier studies that examined the effect of IL on vocabulary learning. First, following earlier guidance on meta-analysis (e.g., Plonsky & Oswald, 2015), a literature search was conducted to identify studies that tested the prediction of the ILH where participants learned new L2 words incidentally. Second, we filtered the identified research reports to determine which studies examined incidental vocabulary learning in an appropriate manner for our meta-analysis. Third, studies were coded for learning conditions (e.g., activity type, frequency, mode, test format and day), a dependent variable (i.e., the reported learning gains), the updated ILH components (need, search, evaluation, sentence-level varied use, composition-level varied use), and the other factors identified as influential (test format, test day, frequency, and mode; Author, XXXXb). Fourth, the included studies were examined (1) to determine types of incidental learning activities that have been used to test the ILH prediction, and (2) to calculate the estimated learning gains for each activity type by using a meta-regression analysis.

4.3.2 Data collection

Literature search. Following earlier recommendations (In'nami & Koizumi, 2010; Plonsky & Oswald, 2015), we searched the following databases to comprehensively identify studies that examined ILH: Linguistics and Language Behavior Abstract (LLBA), Educational Resources Information Centre (ERIC), PsycINFO, ProQuest Global Dissertations, Google Scholar, and VARGA (http://www.lognostics.co.uk/varga). Unpublished research (e.g., doctoral dissertations and master's theses) were also included in order to avoid potential publication biases (Oswald & Plonsky, 2010). We searched studies published from 2001 to April 2019 using a variety of combinations of keywords including involvement load hypothesis, involvement load, task-induced involvement, learning/acquisition/retention,

word/vocabulary, and task. This electronic database search identified 963 reports. Additionally, we conducted a forward citation search with Google Scholar to retrieve studies that cited Laufer and Hulstijn (2001) and included the aforementioned keywords in their titles to identify studies that potentially investigated incidental vocabulary learning and discussed ILH. This forward citation search found 327 more research reports. Consequently, we identified a total of 1290 reports that were potentially eligible to be included in the analysis.

Selection criteria

We created the following six selection criteria to determine which studies to include or exclude from the analysis.

1. The study must have examined vocabulary learning from incidental learning conditions. We followed Hulstijn's (2001) and Laufer and Hulstijn's (2001) definition of incidental vocabulary learning and only included studies when (a) participants were not mentioned about upcoming vocabulary posttests before the treatment and (b) there was no explicit instruction for participants to commit target words to memory. 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, 2012) were excluded. Similarly, studies including deliberate vocabulary learning conditions (e.g., keyword technique, word card learning) were excluded.
2. The study must have investigated incidental vocabulary learning through testing the prediction of the ILH and explicitly coded IL for all learning conditions.
3. The study must have reported enough descriptive statistics to analyze posttest scores (i.e., mean, SD, the number of participants tested).
4. Studies investigating a condition that included multiple language activities were excluded. This is because if participants engage in multiple activities, it is not possible to determine how a single activity contributed to learning gains.
5. We excluded studies when their results were already reported in other publications included in this meta-analysis.

6. The study must have reported their research procedure clearly. First, we excluded studies when learning conditions were not described clearly enough to appropriately code activities. Second, we excluded studies when a study did not report how participants might learn the meanings of target words (e.g., the provision of glosses, dictionary, or guessed from a context). Given that the current study included both published and non-published (e.g., PhD and MA thesis, book chapters) studies, this criterion was important to maintain the quality of included studies.

The abstracts of the research reports identified by the literature search were carefully screened referring to the criteria. We retrieved the full text for 137 reports that appeared to examine vocabulary learning and mention ILH. Forty studies were identified as meeting all of the selection criteria. Additionally, there were 14 studies that were only missing the descriptive statistics. We contacted the authors of these studies and gratefully received information about two studies (Hazrat, 2015; Tang & Treffers-Daller, 2016). In the end, a total of 42 studies (N = 4628) reporting 398 mean posttest scores were included in the analysis. The included studies are comprised of thirty journal articles, four master's theses, three book chapters, two doctoral dissertations, two conference presentations, and one bulletin article (see Appendix A for basic information about the studies).

4.3.3 Coding of Included Studies

First, to identify learning conditions that elicited similar cognitive processing of target words, learning activities were coded for the components of updated ILH (Author, XXXXb). Factors that were also determined to influence incidental vocabulary learning (e.g., test format, mode, and frequency) were also coded and used as covariate to control the effects of factors that that are not directly related to activity type (Author, XXXXb). The reported posttest scores were also coded and standardized to calculate effect sizes (ESs) (see Appendix C and J for the detailed coding scheme).

Because all included studies examined the effect of IL, it may be reasonable to assume that learning conditions with the same combination of updated IL components adequately correspond to the same activity type. This allows objective categorization of

activities that is based on the cognitive processing of target words. We followed Author's (XXXXb) description of the updated ILH components to ensure that the activities were consistently coded according to their learning conditions. Updated ILH components are need, search, evaluation, sentence-level varied use, and composition-level varied use. Therefore, studies' learning conditions were coded as to whether each component was included or not. Because search—which refers to the cognitive conditions where students look for target words or meanings of target words, e.g., by using a dictionary—can be present or absent regardless of activity types, only need and evaluation was used to group learning activities. Search was not used to group activities but used as a covariate to account for its effect on learning gains.

Second, to investigate activities that were used to test the ILH in the studies, each study was coded for the types of activities examined. Activities were first coded using larger categories such as reading, listening, gap-filling, and writing based on the commonly used activity names. Next, to make the labels transparent, each activity type was coded so that it expresses how the updated ILH components are included. Reading and listening were further coded for (i) their reference to target words (e.g., glosses [marginal and glossaries], dictionaries, multiple-choice glosses, and no reference to target words) to distinguish reading involving evaluation (as in reading with multiple-choice glosses) from reading without evaluation (as in glossed reading and reading with a dictionary), and (ii) when a reading or listening activity included need (how need was operationalized i.e., participants had to answer reading comprehension questions that require understanding target words or the material requires the understanding of target words). When the study coded their reading or listening activity as including need, but did not clearly describe how, we trusted the study and coded for the understanding of target words (e.g., reading where target words were important for comprehension). Writing activities were further coded as either sentence writing, composition writing, and summary writing. Similarly, fill-in-the-blanks was coded as either fill-in-the-blanks in a text or fill-in-the-blanks in sentences. The coded activity labels were double-checked to ensure consistency and clarity of labeling (see Appendix K for the final categories and the details of the coding).

4.3.4 Independent Variables

Several factors are not directly related to learning gains but contribute to incidental vocabulary learning. For example, Author (XXXXb) identified test format, test day, frequency, and mode as influencing learning in addition to the ILH components. These factors vary regardless of activity types. When analyzing the reported learning gains to answer RQ2, accounting for these factors may enhance the estimation of learning gains. Therefore, we coded the included studies for independent variables identified by the updated ILH as influencing incidental vocabulary learning (i.e., test format, test day, frequency, and mode).

The first independent variable was test format. How learning is measured greatly influences vocabulary learning gains (Webb, 2005). Therefore, it is important to control for the effects of test format when estimating incidental vocabulary learning gains. Author (XXXXb) identified the optimal grouping of test format in the sampled studies as (a) meaning recall, (b) form recall, (c) recognition (i.e., meaning recognition and form recognition[meaning cue & form cue]), and (d) other test formats (i.e., VKS & use of target words). The present study followed this grouping.

Frequency, mode, and test day were also included as independent variables because these factors were found to influence incidental vocabulary learning in the studies included in this meta-analysis (Author, XXXXb). These variables were used as covariates to control for the variance that was not directly related to activities type. Frequency was coded for the number of times participants used or encountered each target word during an activity. Mode was coded as written when participants engaged in an activity where target words and other language input were provided in a form of written material (e.g., reading, fill-in-the-blanks, writing). Mode was coded as spoken when participants engaged in a spoken activity (listening and speaking; e.g., Jahangard, 2013; Hazrat, 2015) or where target words and other language input were provided in spoken form as well as written form (Snoder, 2017, where participants listened to the text read aloud then read it by themselves). This coding reflects the fact that in activities in spoken mode, students were often provided with written target words, e.g., in a form of a

glossary list (see e.g., Hazrat, 2015, p. 85). Lastly, test day was coded as the number of days between learning and testing.

4.3.5 Coding Procedure and Double Coding

A total of four researchers were involved in the coding process to ensure the consistency and reliability of the coding of the updated ILH components and other independent variables. First, one author and another researcher who had also conducted meta-analyses of vocabulary studies coded three studies separately using the original coding scheme. After confirming that no discrepancy across the two coders was found, potential confusion and ambiguity in the coding scheme was discussed. We revised the coding scheme to enhance its clarity and objectivity so that every study was coded consistently. Subsequently, one author coded all 42 studies carefully, 22 studies (52.4%) were randomly selected and double-coded separately by two other researchers who had also conducted meta-analyses. We calculated the inter-coder reliabilities, Cohen's Kappa coefficient (κ), and confirmed that the coding agreed at a high and acceptable rate (κ = .99 and .98 for each double-coder). We discussed all discrepancies until reaching agreement. Lastly, the first author carefully double-checked all coding to make sure every coding was consistent across studies.

4.3.6 Dependent Variable: Vocabulary Learning Gains

Following earlier meta-analyses of vocabulary studies (Swanborn & de Glopper, 1999; Yanagisawa et al., 2020), we used the proportion of unknown words learned—a.k.a. relative learning gain; see Horst, Cobb, & Meara, 1998; Webb & Chang, 2015—as effect sizes (ESs). The reported posttest scores were standardized by using the following formula.

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

Accordingly, we calculated sampling variances of the ESs from reported SDs converted into proportions by using the escalc function of the metafor package (Viechtbauer, 2010) in R statistical environment (R Core Team, 2020). Each ES was

weighted using the inverse of the sampling variance (Hox, 2010; see Appendix D for the details of formulas used to calculate ES and sampling variance).

4.3.7 Data Analysis

To answer the first research question to determine which types of incidental activities have been used in studies examining the ILH, we first grouped activities according to their ILH components in order to categorize activities that included similar cognitive processing. The activities were grouped into one of the following five combinations of need and evaluation components: (1) when there is no need or no evaluation, (2) when there is need without evaluation, sentence-level varied use, or composition-level varied use, (3) when there is need and evaluation, (4) when there is need and sentence-level varied use, and (5) when there is need and composition-level varied use. We followed this grouping scheme and categorized activities.

To answer the second research question, we determined the estimated mean learning gains for the different activity types. As in the earlier meta-analysis (Author, XXXXa, XXXb; de Glopper & Swanborn, 1999; Yanagisawa, Webb, & Uchihara, 2020), reported posttest scores were converted into the proportion of unknown words learned and analyzed by using a three-level meta-regression model (Cheung, 2014). First, the mean of the overall incidental learning gain was calculated using an intercept only model. Second, we calculated the mean learning gains for each activity type. Following the activity grouping scheme of Research Question 1, we created a new categorical variable, activity type, which had five levels indicating each combination of need and evaluation components.

The other predictor variables, which can vary regardless of activity type (i.e., search, frequency, mode, test format, and test day), were used as covariates to control for their effects. First, the estimated mean learning gains for different types of activities were calculated by using no-intercept models (Yanagisawa, Webb, & Uchihara, 2020). Second, the difference in mean learning gains across activity types was compared by changing the reference level of the activity type variable (de Vos et al., 2018).

Lastly, to reflect the variance in learning gains across studies, we calculated the predictor interval for each activity type. Predictor intervals provide ranges of estimated learning gains, with which one can predict the extent to which L2 students will learn new unknown words based on the types of activity and learning conditions they engage in.

We adopted a three-level meta-regression model (Cheung, 2014; Lee et al., 2018) to analyze ESs expressing the proportion of unknown words learned (de Glopper & Swanborn, 1999; Yanagisawa et al. 2020). One advantage of three-level meta-regression models over common meta-regression models is that it accounts for different sources of variance related to ESs (i.e., sampling variance, the variances within each study and across studies), so it enables more reliable analyses of learning gains from different conditions examined within each study (e.g., Yanagisawa et al. 2020). The majority of studies reported multiple posttest scores that were not independent due to sampling error (e.g., learning gains of the same participants were tested multiple times or with different measurements), thus potentially increasing a Type I errors. To cope with this bias, cluster robust variance estimation (Hedges et al., 2010) with small sample adjustments (Tipton, 2015; Tipton & Pustejovsky, 2015) were used when assessing the statistical significance of predictor variables.

All analyses were conducted in R (R Core Team, 2020). The metafor package's rma.mv function (Viechtbauer, 2010) was used to fit three-level meta-regression models with maximum likelihood estimation. Three different sources of variance were accounted for: level 1, sampling variance of the effect sizes; level 2, variance between effect sizes from the same study (within-study variance); and level 3, variance across studies (between-study variance). The ClubSandwich package (Pustejovsky, 2018) was used to calculate *p*-values and confidence intervals (CIs) based on the robust variance estimation. The learning gains on the immediate posttest and the delayed posttest were analyzed separately to reveal estimated learning gains for activities at two retention intervals: immediately after the treatment and on delayed posttests.

4.4 Results

In answer to the first research question, learning activities were grouped according to the presence of components of the updated ILH (need, evaluation, sentence-level varied use, composition level varied use) that have been found to contribute to incidental vocabulary learning. Table 1 shows the activities according to their updated ILH components (see also the completed coding scheme that is publicly available via OSF). First, learning conditions including none of the components were comprised of three activities that accounted for a total of 20 ESs: glossed reading (12 ESs), listening with a list of target words (6 ESs), and reading without reference to target words (2 ESs). These activities refer to meaning-focused input (MFI) activities (i.e., reading or listening) focused on comprehension and there was no clear need for participants to understand the meanings of target words.

Second, conditions including only need were composed of five activities that accounted for a total of 88 ESs: glossed reading with comprehension questions requiring the understanding of target word (55 ESs), glossed reading where target words were important for comprehension (18 ESs), listening with a list of target words and comprehension questions requiring the understanding of target word (10 ESs), reading in which target words were important for comprehension and dictionaries were provided (3 ESs), and reading with the support of dictionaries plus comprehension questions requiring the understanding of target word (2 ESs). This group also corresponds to listening and reading MFI activities. However, in contrast to the first activity type (MFI activities), these activities required participants to answer comprehension questions that required the understanding of the target words (e.g., Hulstijn & Laufer, 2001). In contrast to the evaluation activity type, these activities did not require participants to evaluate multiple meanings of target words.

Third, learning conditions including need and evaluation comprised of eight activities that accounted for a total of 161 ESs: fill-in-the-blanks in passages (91 ESs), reading with multiple-choice glosses (26 ESs), fill-in-the-blanks in sentences (14 ESs), multiple-choice questions (15 ESs), translation (5 ESs), matching (4 ESs), reading with

dictionaries (multiple meanings were presented for each target word and participants needed to determine the meaning that fit the context: 4 ESs), and sentence-combinations, where participants combine segments of a sentence to regenerate the sentence (2 ESs). These activities can be referred as evaluation activities as they include the comparison of meanings or words related to target words.

Fourth, conditions with need and sentence-level varied use included three activities that accounted for a total of 72 ESs: sentence writing (62 ESs), graphic organizers involving sentence-production (6 ESs), and oral sentence-production (4 ESs). In these activities, students were asked to use a target word in a sentence, or to verbally generate a sentence including a target word.

Lastly, conditions with need and composition-level varied use were comprised of three activities that accounted for a total of 57 ESs: composition writing (47 ESs), retelling (8 ESs), and summary writing (2 ESs). These activities required participants to use a set of target words to create a cohesive written text.

Based on the included activities, five different combinations of learning conditions were labeled as (i) meaning-focused input (MFI) activities, (ii) MFI with need for comprehension of target words, (iii) evaluation activities, (iv) sentence-level varied use activities, and (vi) composition-level varied use activities. This grouping was used to create a categorical predictor variable indicating the type of activities in order to answer the second research question.

Table 1: Incidental vocabulary learning activities classified according to the updated ILH features that contribute to learning

Combinations of the updated ILH components				
Need	Evaluation	Sentence-level varied use	Composition-level varied use	Activity type
0	0	0	0	MFI
1	0	0	0	MFI with need for comprehension of target words
1	1	0	0	Evaluation activities
1	0	1	0	Sentence-level varied use activities
1	0	0	1	Composition-level varied use activities

Note. MFI = meaning-focused input activity.

To answer the second research question, three-level meta-regression models were fitted with data to estimate mean learning gains that occur through completing the different vocabulary learning activities. First, the intercept only model was fitted to examine whether the amount of variance in ESs were related to different sources, i.e., differences in learning gains within each study and those across different studies. The variance distribution was examined by calculating I^2 indices using the dematar package (Harrer, Cuijpers, Furukawa, & Ebert, 2019; see also Cheung, 2014). On immediate posttests, I^2 indices showed that 0.50% of the variance was attributed to the sampling variance, 50.58% to the within-study level variance—which reflects learning gain differences in each study—, and 48.92% to the between-study level variance—which reflects the difference across different studies. On the delayed posttests, I^2 indices showed that 0.43% of the variance was attributed to the sampling variance, 61.93% to the within-study level, and 37.64% to the between-study level. This suggests that the sampling error of individual studies was quite small and that the differences in learning gains on immediate posttests were mainly due to the differences between studies (e.g., using different target words and participant groups) and the differences within studies (e.g., using different activities, test formats, and test timings). The fact that about 37%-50% of the variance on posttests was attributed to between-study level variance suggests that learning gains differed greatly from study to study, pointing to the possibility that the characteristics of participant groups and target words—which were different from study to study and were not considered in the current analysis—may have impacted the reported learning gains considerably. The results of the intercept only model revealed that in all activities combined 43.9% and 32.7% of unknown words were learned on the immediate (b = 0.439, 95% CI [0.378, 0.500], $p < .001$) and the delayed posttests (b = 0.327, 95% CI [0.261, 0.393], $p < .001$), respectively. To obtain the estimated mean learning gains for each activity type, the model with the categorical variable indicating the activity type (i.e., MFI, MFI with need for comprehension of target words, evaluation, sentence-level varied use, and composition-level varied use) and other variables (i.e., test format, test timing, frequency, search, and mode) was fitted. In order to make the interpretation easier, these covariate variables were centered or set for their reference levels. Test format had four levels (i.e., meaning recall, form recall, recognition, and

other test formats) and meaning recall was set as the reference level. Thus, learning gains were estimated for when a meaning recall test (e.g., L2 to L1 translation test) was used. Frequency was centered at 1 to indicate 1 encounter/use of each target word as the reference level. Similarly, test day for the delayed posttests was centered at 14 days after learning, thus indicating the estimated learning gains were for when learning is measured 2 weeks after the learning session. Mode had two levels (written, spoken) and written was set as the reference level.

Table 2 presents the estimated mean learning gains on immediate and delayed posttests, separately (see Appendix K for the details of the results including all predictor variables). The results of a no-intercept model on the immediate posttests showed that the effectiveness of activity type was in the following order (estimated learning gains are presented in parentheses following each activity type): composition-level varied use (61.0%), sentence-level varied use (53.0%), evaluation activities (46.2%), MFI with need (37.8%), and MFI (16.5%). A Wald-test showed that the activity type significantly influenced learning gains, $F(13.1) = 9.88$, $p < .001$. Subsequent multiple comparisons detected statistical differences across all activity types at an alpha level of $p < .05$, except between composition-level varied use and sentence-level varied use ($p = .100$). Composition-level varied use led to significantly greater learning gains than MFI ($p = .002$), MFI with need for comprehension of target words ($p < .001$), and evaluation ($p = .003$). Sentence-level varied use also led to greater learning than MFI ($p = .001$), MFI with need ($p < .001$), and evaluation ($p = .007$). Evaluation outperformed MFI ($p = .007$) and MFI with need for comprehension of target words ($p = .001$). MFI with need for comprehension of target words led to significantly greater learning than MFI ($p = .028$).

Table 2: Estimated Learning Gains (the Proportion of Target Words Learned)

	Immediate Posttest					Delayed Posttest				
				CI					CI	
Activity Type	k	n	Mean ES	Lower	Upper	k	n	Mean ES	Lower	Upper
MFI	7	10	0.165	0.010	0.320	6	10	0.126	0.006	0.245
MFI with Need for comprehension of target words	21	44	0.378	0.296	0.459	21	44	0.266	0.195	0.337
Evaluation	30	84	0.462	0.384	0.540	30	77	0.356	0.286	0.427
Sentence-level Varied use	20	36	0.530	0.453	0.608	20	36	0.379	0.313	0.445
Composition-level Varied use	13	33	0.610	0.490	0.731	13	24	0.474	0.394	0.553

Note. k = number of studies. n = number of ESs. CI = 95% confidence interval adjusted with robust variance estimation. MFI = meaning-focused input (i.e., reading and listening activities). The total number of studies = 42. Total number of ESs = 398. Mean learning gains were estimated for when a meaning recall test was used, mode was written, search was not included, and frequency was 1. For the immediate posttest, learning gains were estimated for when measured on the same day as learning, and for the delayed posttest, learning gains were when learning was measured 14 days after the learning session.

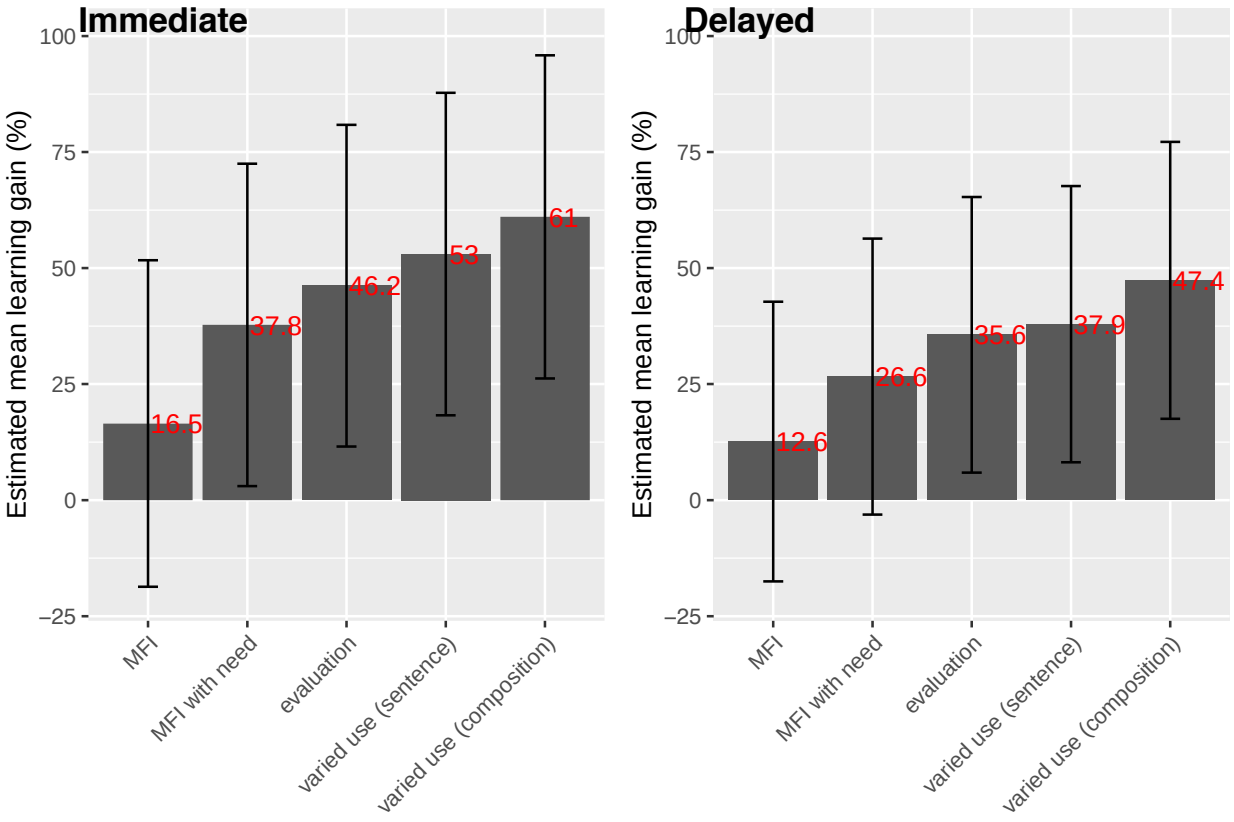
The results of a no-intercept model on the delayed posttest revealed that the effectiveness of activity type was in the same order as the results on the immediate test (estimated learning gains are presented in parentheses following each activity type); composition-level varied use (47.4%) led to the greatest learning gains, followed by sentence-level varied use (37.9%), evaluation activities (35.6%), MFI with need (26.6%), and MFI (12.6%). A Wald-test showed that the activity type significantly impacted learning gains, $F(13.3) = 16.4$, $p < .001$. Subsequent multiple comparisons found statistical differences across all activity types at an alpha level of $p < .05$, except between sentence-level varied use and evaluation activities ($p = .274$). Composition-level varied use led to greater learning gains than MFI ($p < .001$), MFI with need for comprehension of target words ($p < .001$), evaluation ($p = .003$), and sentence-level varied use ($p < .003$). Sentence-level varied use led to greater learning than MFI ($p = .002$), and MFI with need for comprehension of target words ($p < .001$). Evaluation outperform MFI ($p = .003$) and MFI with need for comprehension of target words ($p = .001$). MFI with need for comprehension of target words led to significantly greater learning than MFI ($p = .028$).

In order to extend the results of the estimated mean learning gains for predicting future learning gains, prediction intervals (PIs) were calculated for each estimated mean learning gain using the predict function of the metafor package. The values were converted into percentages for the sake of interpretability. Figure 1 shows the estimated percentage of unknown words that will be learned for each activity type and their 90% PI. The 90% PIs indicate the range in which the future observation will fall with a probability of 90%. On the immediate posttests, the estimated learning gains and their calculated 90% PIs were 61.0%, PI [26.2, 95.5] for composition-level varied use activities; 53.0%, PI [18.3, 87.8] for sentence-level varied use activities; 46.2%, PI [11.5, 80.9] for evaluation activities; 37.8%, PI [3.0, 72.5] for MFI with need for comprehension of target words; 16.5%, PI [-18.7, 51.7] for MFI.

On the delayed posttests, the estimated learning gains and their calculated 90% PIs were 47.4%, PI [17.5, 77.2] for composition-level varied use activities; 37.9%, PI

[8.1, 67.7] for sentence-level varied use activities; 35.6%, PI [5.9, 65.3] for evaluation activities; 26.6%, PI [-3.1, 56.4] for MFI with need; 12.6%, PI [-17.5, 42.8] for MFI.

Figure 1: The Estimated Mean Learning Gains for Different Types of Activities



Note. Error bars indicate 90% prediction intervals. MFI = meaning-focused input (i.e., reading and listening activities). need = need for comprehension of target words, i.e., a condition where participants clearly have to understand target words. Learning gains were predicted for when a meaning recall test was used, mode was written, search was not included, and frequency was 1. Immediate Test = when learning is measured on the same day as the learning session. Delayed Test = when learning is measured on 14 days after the learning session.

4.5 Discussion

In answer to the first research question, learning conditions that have been examined to test the ILH prediction were grouped into five activity types according to the components of the updated ILH: MFI, MFI with need for comprehension of target words, evaluation, sentence-level varied use, and composition-level varied use.

Two types of MFI activities were identified: (a) MFI and (b) MFI with need for comprehension of target words. MFI was comprised of activities in which there was no need for participants to know the target words; the aim of the activities was comprehension—i.e., glossed reading (e.g., Tang & Treffers-Daller, 2016; Yang et al., 2017), reading without reference to target words (i.e., Beal, 2007), and listening with a glossary (Jing & Jianbin, 2009; Maleki, 2012). MFI with need for comprehension of target words differs from MFI in that it was necessary for participants to know the target words in order to complete the activities. Examples of activities in this category included reading to answer comprehension questions that require the understanding of target words and reading where target words were important for comprehension. The main difference between MFI and MFI with need for comprehension of target words is that the latter required participants to understand the meanings of target words by either asking them to answer comprehension questions that were related to the target words (e.g., Hulstijn & Laufer, 2001; Kim, 2008) or by using a text where knowledge of the target words was important for comprehending the text (e.g., Cao, 2013; Rott, 2012). Analysis comparing these two activity types may reveal how useful it is to implement the necessity of target words in MFI activities to facilitate vocabulary learning.

Evaluation corresponded to activities where participants compared the meanings of target words or forms of target words. This activity type included fill-in-the-blanks, reading with multiple-choice glosses, multiple-choice questions, and matching. Three activities (i.e., reading with multiple-choice glosses, reading with dictionaries where multiple meanings for each target word were presented, translation from L2 to L1) require an explanation of how they fit into the evaluation category. In reading with multiple-choice glosses, participants not only needed knowledge of the target words in

order to complete the task (MFI with need for comprehension of target words), but they also had to select the meanings that fit the contexts best from among several options (e.g., Martínez-Fernández, 2008). Reading with dictionaries where multiple meanings for each target word were presented was an activity where participants were provided with multiple meanings in a dictionary for target items and had to select the appropriate meanings (e.g., Yaqubi, Rayati, & Gorgi, 2010). In reading with multiple-choice glosses and reading with dictionaries where multiple meanings for each target word were presented, it is the need to determine the correct option from several choices that involves evaluation. Translation from L2 to L1 involved participants translating L2 sentences into L1 sentences (e.g., Bao, 2015). Laufer and Girsai (2008) coded receptive translation as including *moderate evaluation* by explaining that receptive translation requires participants to evaluate the multiple translational alternatives (i.e., different L1 translations of a word) in order to write the appropriate words to fit the context (p. 712). Taken together, it is the determination of the appropriate L2 forms or meanings that signals that activities were coded as evaluation.

Activity types including varied use were comprised of sentence-level varied use and composition-level varied use based on whether target words were used individually or collectively. Sentence-level varied use activities included three activities: sentence writing, graphic organizers involving sentence-production, and oral sentence-production. These activities asked participants to use a single target word in a sentence in written or spoken mode. In contrast, composition-level varied use required participants to use a set of target words to create a coherent written text in activities such as composition writing, retelling, and summary writing.

The creation of the activity types is useful because earlier studies of the ILH reveal that categorizing activities according to ILH components is challenging (Authors, XXXXa). The fact that researchers conducting studies of the ILH may have trouble coding activities according to their motivational and cognitive components suggests that teachers may also have difficulty interpreting the research and applying it to their teaching practice. In contrast, when activities are grouped based on the processes that contribute to learning target words, the application of research findings may be more

easily applied to the selection of activities for teaching and learning. This may enable materials designers, teachers, and learners who are not familiar with research to apply it to teaching and learning. This might also be a useful step to making applied linguistics research and L2 learning research more easily incorporated into L2 language teaching and pedagogy. Within studies of applied linguistics, we need to have very transparent activity labels that can be applied to pedagogy. Similarly, categorizing learning conditions based on the included factor contributing to learning could also be applied to summarizing other research findings such as activities that aim to promote grammatical knowledge (e.g., the acquisition of past tense) or communicative competence (e.g., reading comprehension ability).

In answer to the second research question, the findings indicate that incidental vocabulary learning gains differed significantly among activity types. Figure 1 shows estimated mean learning gains (i.e., the percentage of unknown words learned) and their 90% PIs for different activity types. The estimated values were calculated for conditions where learning was measured with a meaning recall test, mode was written, search was not included, and frequency was 1. The estimations were also based on testing on the same day as learning in immediate posttests and two weeks after the learning session in delayed posttests. The types of activities according to their estimated mean percentage of unknown words learned from the most effective to the least effective were: (a) composition-level varied use (61.0% and 47.4% of target items were revealed to be learned on immediate and delayed posttests, respectively), (b) sentence-level varied use (53.0% and 37.9%), (c) evaluation activities (46.2% and 35.6%), (d) MFI with need for comprehension of target words (37.8% and 26.6%), (e) MFI (16.5% and 12.6%) in that order.

The order of activity effectiveness was also supported by the subsequent multiple comparisons. There were statistically significant differences across all activity types, except between composition-level varied use and sentence-level varied use ($p = .100$) on the immediate posttest and between sentence-level varied use and evaluation ($p = .274$) on the delayed posttest. This means that when conducting research comparing these activity types, there may typically be a significant difference between the activities

except that the advantages of composition-level over sentence-level varied use, and sentence-level varied use over evaluation might not be detected. However, given the larger mean learning gains of composition-level varied use over sentence-level varied use and those of sentence-level varied use over evaluation, the former activities are more likely to yield greater learning gains; thus, they should be recommended over the latter activities.

The result that MFI with need for comprehension of target words led to greater learning gains than MFI highlights the importance of designing reading and listening activities in which students need to understand target words. The results are supported by earlier studies indicating that students tend to ignore unknown L2 words (Ender, 2016; Hulstijn et al., 1996) even when glosses are provided (Boers et al., 2017; Warren et al., 2018). Additionally, Jin and Webb (2019) found that students learn more words from teacher talk when they took notes on unknown target words and their meanings. This alludes to the possibility that vocabulary learning from MFI can be enhanced if learners engage with unknown words in some way. The majority of the studies in the current meta-analysis implemented comprehension questions to elicit need for comprehension of target words, thus it may be useful for future research to explore the effects of different techniques requiring students to process target words such as note-taking.

The ranking of activities showed a clear advantage of productive activities (writing and speaking) over receptive activities (reading and listening). One plausible explanation for this is that using target words in an original context may induce more elaborated cognitive processes such as thinking of how words should be used with other words in a grammatically and semantically appropriate manner with acceptable collocations (Laufer and Hulstijn, 2001). In contrast, receptive activities tend to only require learners to attend to the form-meaning connections of words (Kaivanpanah & Miri, 2018). The finding is in line with the *output hypothesis* (Swain, 1985) and the *generation effect* (Slamecka & Graf, 1978), both of which suggest that using language productively plays a critical role in learning. Vocabulary research has also frequently documented the advantage of productive activities over receptive activities (Huang et al., 2012; Webb, 2005, 2009; but also see Shintani, 2011).

The grouping of activity types also provides language teachers with an approximation of the extent to which students learn new words through completing different activities. While the estimated mean percentage of learning gains for different activity types may best illustrate the ranking of activity types, it does not necessarily mean that L2 students always learn as much vocabulary as the mean scores because learning gains widely fluctuate based on a variety of factors such as the characteristics of participants and target words in addition to the type of activities they engage in. Taking advantage of a meta-regression model, where different sources of the variance in ESs were accounted for, the PIs of the mean percentage of learning gains was calculated. The 90% PI of the mean estimated learning gains on the immediate posttest were 26.2% to 95.5% for composition-level varied use, 18.3% to 87.8% for sentence-level varied use, 11.5% to 80.9% for evaluation, 3.0% to 72.5% for MFI with need for comprehension of target words, and -18.7% to 51.7% for MFI. Because the definition of estimated percentage of words learned is never a negative value, the lower PI for MFI is interpreted as 0. Looking at composition-level varied use as an example, the PI indicates that we are 90% confident that future mean percentages of unknown words will be learned will fall within 26.2% to 95.5%. On the delayed posttest, the 90% PI of the mean estimated learning gains were 17.5% to 77.2% for composition-level varied use, 8.1% to 67.7% for sentence-level varied use, 5.9% to 65.3% for evaluation, -3.1% to 56.4% for MFI with need for comprehension of target words, and -17.5% to 42.8% for MFI.

The calculated PIs reveal that the learning gains differ greatly from study to study. Although the meta-regression models accounted for the effects of learning conditions, test formats, and other influential variables (i.e., frequency, mode, search component, test day), the results also demonstrate that incidental learning conditions are also greatly influenced by other factors that were not considered in this study. Given that about 37-50% of the variance in ESs were due to the variance at the between-study level—which reflects differences across studies—as indicated by I^2 values, the wide range of PIs may reflect the fact that vocabulary learning gains fluctuate among learners, contexts, and words. This corresponds to the reality of vocabulary learning, in which there is no guarantee that something will be learned through completing a single activity or that something will be learned to the same degree by all learners in all contexts.

Although one might question the value of the wide range of PIs, they are still useful for language teachers as the PIs provide a rough idea of the minimum and maximum percentage of words that will be learned. For example, on the immediate posttest, the upper band of PI for composition-level varies use approaches 100%, suggesting that students will potentially learn almost all target words through engaging in this activity type. On the other hand, the upper PI for MFI activities where there is no clear need to understand target words is about 50%, indicating that there is only a small chance that learners learn more than half of the target words.

4.5.1 Limitations and Future Directions

The present study provided a first attempt to create a predictive model to estimate the amount of vocabulary learning based on learning conditions. The results provide a useful ranking of activity types for vocabulary learning, as well as estimates of potential learning gains through completing the different tasks. The results also showed that the calculated PIs were relatively wide. It may be useful for future studies to try enhancing the prediction by considering other factors such as the characteristics of students (e.g., prior vocabulary knowledge, Webb and Chang, 2015) and target words (e.g., number of letters, pronounceability, imageability, concreteness, Ellis and Beaton, 1993) in addition to learning conditions and measurement related variables.

It would also be useful for future studies is to compare the relative effectiveness of the activities according to their rankings. The results showed that composition-level varied use activities led to the greatest learning gains, followed by sentence-level varied use, evaluation activities, MFI with need for comprehension of target words, and MFI. Future empirical studies could compare these activities to examine whether the efficacy ranking was as predicted.

One ambiguity found in the ILH studies relates to the description of activities. Some studies did not clearly state how they ensured that participants had to understand target words. Similarly, several studies did not clearly state how the evaluation component differed between MFI with need and evaluation. Future studies are encouraged to clearly describe how the IL components are included within different

activities. Making the research materials publicly available may also enhance the transparency of the research design.

It is also important to note that although the results showed that MFI led to the smallest learning gains, this does not mean that activities such as extensive reading, listening, and viewing should be abandoned. The majority of the included studies only looked at one learning session—usually reading one short text where target words occurred only once. Including all studies of incidental vocabulary learning in MFI was beyond the scope of the present study. However, MFI activities where learners repeatedly encounter target words in a variety of contexts over a longer period would likely lead to greater vocabulary learning, as well as the development of other aspects of vocabulary knowledge (e.g., collocation, word parts, association) beyond form-meaning connection (Webb & Chang, 2015). Although several studies followed vocabulary development through engaging in extensive reading (e.g., Webb & Chang, 2015), no study has compared different types of incidental vocabulary learning activities in a longitudinal design. Further research in this area is warranted.

Finally, while the current study focused on the effects of language activities on vocabulary learning, it may be useful to meta-analyze studies that examined the effect of other types of activities such as those focused on intentional vocabulary learning, or the learning of grammatical knowledge and skills. Although studies have meta-analyzed the effects of grammar instruction (e.g., Shintani, Li, & Ellis, 2013; Shintani, 2015), it is still unclear how different language activities contribute to grammar acquisition. Further meta-analyses may deepen our understanding of how L2 knowledge develops by engaging in other activities and provide useful pedagogical implications as to how L2 learning can be optimized.

4.6 Conclusion

The present study meta-analyzed the studies testing the ILH to provide an overview of the different incidental vocabulary learning conditions that have been examined in studies of the ILH and obtain the estimated vocabulary learning gains that

occur across those activities. The results showed that many learning conditions were adopted to test the ILH and these conditions were classified into five activity types according to the factors within activities that contribute to learning. The activity types were MFI, MFI with need for comprehension of target words, evaluation, sentence-level varied use, and composition-level varied use. The results showed that composition-level varied use activities led to the greatest learning gains, followed by sentence-level varied use, evaluation activities, MFI with need for comprehension of target words, and MFI. Additionally, the estimated learning gains and their predictive intervals were calculated for each activity type. Thus, one can easily estimate the relative efficacy of activities on vocabulary learning and the extent to which L2 students learn new words based on the provided activity types.

4.7 References

The full reference list of the studies included in the meta-analysis is available in Appendix G.

Authors (XXXXa) refers to the current thesis's Study 1

Authors (XXXXb) refers to the current thesis's Study 2

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

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

Boers, F., Warren, P., Grimshaw, G., & Siyanova-Chanturia, A. (2017). On the benefits of multimodal annotations for vocabulary uptake from reading. *Computer Assisted Language Learning*, *30*(7), 709–725. https://doi.org/10.1080/09588221.2017.1356335

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag. //www.springer.com/la/book/9780387953649

Cao, Z. (2013). The effects of tasks on the learning of lexical bundles by Chinese EFL learners. *Theory and Practice in Language Studies*, *3*(6), 957–962. https://doi.org/10.4304/tpls.3.6.957-962

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

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*(2), 227–252. https://doi.org/10.1177/1362168811431377

Ender, A. (2016). Implicit and Explicit Cognitive Processes in Incidental Vocabulary Acquisition. *Applied Linguistics*, *37*(4), 536–560. https://doi.org/10.1093/applin/amu051

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2019). Doing meta-analysis in R: A hands-on guide. *PROTECT Lab Erlangen*.

Hazrat, M. (2015). The effects of task type and task involvement load on vocabulary learning. *Waikato Journal of Education*, *20*(2), 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*(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*(2), 207–223.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2. ed). Routledge, Taylor & Francis.

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*(4), 544–557. https://doi.org/10.1111/j.1540-4781.2012.01394.x

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*(3), 327–339. https://doi.org/10.2307/329439

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 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*(1), 169–184. https://doi.org/10.5054/tq.2010.215253

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.

Jing, L., & Jianbin, H. (2009). An empirical study of the involvement load hypothesis in incidental vocabulary acquisition in EFL listening. *Polyglossia*, *16*, 1–11.

Kaivanpanah, S., & Miri, M. (2018). Inspecting task-induced involvement from the perspective of sociocultural theory. *Journal of Teaching Language Skills*, 37(1), 159–192. https://doi.org/10.22099/jtls.2019.30652.2569

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, *12*(3), 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*(2), 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

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*(4), 567–587.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694–716. https://doi.org/10.1093/applin/amn018

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

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

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*(1), 24–38.

Maleki, N. A. (2012). The effect of the involvement load hypothesis on improving Iranian EFL learners' incidental vocabulary acquisition in listening

comprehension classes. *Australian Journal of Basic and Applied Sciences*, *6*(9), 119–128.

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.

Nguyen, C.-D., & Boers, F. (2018). The effect of content retelling on vocabulary uptake from a TED talk. *TESOL Quarterly*. https://doi.org/10.1002/tesq.441

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

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition*, 1–24. https://doi.org/10.1017/S0272263119000020

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

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. (2020). *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*(4), 689–717.

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

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.

Shintani, Natsuko. (2011). A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Language Teaching Research*, *15*(2), 137–158. https://doi.org/10.1177/1362168810388692

Shintani, N. (2015). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics*, *36*(3), 306–325. https://doi.org/10.1093/applin/amu067

Shintani, Natsuko, Li, S., & Ellis, R. (2013). Comprehension-Based Versus Production-Based Grammar Instruction: A Meta-Analysis of Comparative Studies: Meta-Analysis of CBI and PBI. *Language Learning*, *63*(2), 296–329. https://doi.org/10.1111/lang.12001

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592.

Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of involvement load, spacing, and intentionality. *TESL Canada Journal*, *34*(3), 140–164. https://doi.org/10.18806/tesl.v34i3.1277

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and Practice in Applied Linguistics* (Vol. 2, pp. 125–144). Oxford University Press.

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

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. http://link.springer.com/content/pdf/10.1057/9781137449788.pdf#page=140

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

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

Tsubaki, M. (2012). *Vocabulary learning with graphic organizers in the EFL environment: Inquiry into the involvement load hypothesis* [Unpublished doctoral dissertation]. Temple University.

van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*(3), 609–624. https://doi.org/10.1016/j.system.2013.07.012

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

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 1–17.

Warren, P., Boers, F., Grimshaw, G., & Siyanova-Chanturia, A. (2018). The effect of gloss type on learners' intake of new words during reading: Evidence from eye-

tracking. *Studies in Second Language Acquisition*, 1–24. https://doi.org/10.1017/S0272263118000177

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52.

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*(3), 360–376. https://doi.org/10.1177/0033688209343854

Webb, S. (2020). Incidental Vocabulary Learning. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 225–239). Routledge.

Webb, S., & Chang, A. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, *37*(4), 651–675. https://doi.org/10.1017/S0272263114000606

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*(1), 13–40.

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

Yaqubi, B., Rayati, R. A., & Gorgi, N. A. (2010). The involvement load hypothesis and vocabulary learning: The effect of task types and involvement index on L2 vocabulary acquisition. *Journal of Teaching Language Skills*, *29*(1), 145–163. https://doi.org/10.22099/jtls.2012.404

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*(1), 54–75. https://doi.org/10.1177/1362168816652418

Chapter 5

5 Discussion and Conclusion

This chapter will review the main findings of Studies 1-3 and discuss the implications of the studies as a whole. Finally, limitations of the three studies and directions for further research will be discussed.

5.1 Review of the Findings

5.1.1 Study 1

Study 1 (Chapter 2) meta-analyzed 398 reported posttest scores from 42 empirical studies (N = 4628) in order to explore (1) the overall predictive ability of the ILH, (2) the relative effects of different components of the ILH (need, search, evaluation), and (3) the influence of potential factors moderating learning (e.g., time on task, frequency, test format). Results showed that the ILH was significantly predictive of learning and the clear positive correlation between IL of tasks and learning gains was observed. The results also suggested that the predictive ability of the ILH is not so high. ILH explained 15.0% and 5.1% of the variance in effect sizes on immediate and delayed posttests, respectively.

Each component of the ILH was found to contribute differently to learning. The evaluation component contributed to the greatest amount of learning, followed by need. Interestingly, search was not found to contribute to learning. Pedagogically, the results indicated that evaluation, especially strong evaluation was the component that most contributed to learning, highlighting the value of productive activities involving strong evaluation such as writing and speaking where learners use target words in original sentences or compositions. In contrast, given the fact that search was not found to positively affect learning, language teachers and learners might not need to try implementing a condition in learning where learners search for information e.g., by consulting a dictionary.

Moderator analyses revealed that Involvement Load (IL) had a greater impact on learning than time on task. Although time on task was positively correlated with learning gains, this trend disappeared when IL was controlled. Additional analysis found a positive correlation between IL and time on task, suggesting that the effect of tasks taking longer is mainly due to a greater IL (Hulstijn & Laufer, 2001). This indicates that engaging in a longer task may not necessarily lead to greater learning and that the IL of the task better explains vocabulary learning.

Lastly, the results indicated that frequency positively contributed to learning and no interaction between frequency and IL was found. Pedagogically, this highlights the value of activities where learners encounter or use the same target words. Additionally, the results also showed that the frequency effect was not found on delayed posttests, pointing to the advantage of IL of tasks over frequency when focusing on long term retention.

5.1.2 Study 2

Study 2 (Chapter 3) aimed to update the ILH to enhance its accuracy in predicting incidental vocabulary learning. The information-theoretic approach was adopted to determine the optimal statistical model (i.e., a set of predictor variables) that best predicts learning gains. Following earlier research findings, we investigated whether the prediction of the ILH improved by (i) examining the influence of each level of individual ILH components (need, search, and evaluation), (ii) adopting optimal test format grouping and best operationalization of the ILH components, and (iii) including other empirically motivated variables.

The results revealed that the main factors contributing to the prediction of learning gains were (a) need, (b) evaluation, (c) sentence-level varied use, and (d) composition-level varied use. As discussed previously (Laufer & Hulstijn, 2001; Kim, 2008), examining the contributions of the IL components on their own, rather than the combined IL components as a whole, significantly enhanced the prediction. Furthermore, modifying the evaluation component by distinguishing between different types of *strong evaluation* (relabeled as sentence-level varied use and composition-level varied use) also

improved the accuracy of the prediction. Categorizing test formats into receptive recall, productive recall, and recognition, VKS, and use was found to be the test format grouping that indicated the highest model fit.

The analysis of the other empirically motivated variables indicated that (1) spoken activities (listening and speaking) tended to lead to lower learning gains than written activities (e.g., reading, writing, gap-filling) on the immediate posttests, and (2) frequency positively contributed to learning on the immediate posttests. However, these factors were not useful predictors on the delayed posttest. Additionally, it was found that including search in an activity potentially hinders learning. When search was present, learning retention measured on delayed posttests decreased.

The resulting statistical models showed greater predictive ability, as indicated by the larger explained variance compared to the original ILH. This suggests that the updated ILH predicts 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. Based on the models, we created an Involvement Load (IL) formula. Using this formula, one can calculate the updated IL of activities to more accurately predict their relative effectiveness for incidental vocabulary learning. The prediction based on the IL formula was proposed as an updated ILH.

5.1.3 Study 3

Study 3 (Chapter 4) used a meta-analytic approach to (a) overview the different incidental vocabulary learning conditions that have been examined in studies of the ILH, and (b) obtain the estimated vocabulary learning gains occurring across different activity types.

Learning conditions examined by studies testing the ILH were classified into five activity types according to the factors within the activities that were identified as contributing to the prediction by Study 2. The identified activity types were meaning-focused input (MFI; e.g., reading and listening), MFI with need for comprehension of target words (e.g., reading and listening where learners clearly had to understand each

target word), evaluation (e.g., fill-in-the-blanks), sentence-level varied use (e.g., sentence-writing), and composition-level varied use (e.g., composition-writing).

The reported posttest scores were standardized as effect sizes of the proportion of unknown words learned and analyzed with a meta-regression model. We calculated the estimated learning gains and their predictive intervals for each activity type. The results showed the estimated mean learning gains were highest for composition-level varied use activities (61.0% and 47.4% of target items were revealed to be learned on immediate and delayed posttests, respectively), followed by sentence-level varied use (53.0% and 37.9%), (c) evaluation activities (46.2% and 35.6%), MFI with need for comprehension of target words (37.8% and 26.6%), and MFI (16.5% and 12.6%) in that order. Additionally, predictive intervals of the mean percentage of learning gains were calculated to provide language teachers with a rough idea of the minimum and maximum percentage of words that will be learned through each of the activity types. This study also summarized learning gains by categorizing learning conditions into different activity types that involve the same cognitive processes, hopefully providing more transparent findings that can be easily applied to pedagogy.

5.2 Overall Discussion

Now let us consider the theoretical and pedagogical implication of this thesis when taken as a whole.

5.2.1 Theoretical Implications

The ILH is the most extensively discussed theoretical framework in L2 vocabulary research. Since proposed (Laufer & Hulstijn, 2001), many studies examined whether the ILH accurately predicts the relative effects of activities on vocabulary learning. Although several studies pointed to potential directions to update the ILH (e.g., Folse, 2006; Kim, 2008; Zou, 2017), it was difficult to confirm how the ILH should be revised because research findings were inconsistent. For example, while Zou (2017) found that composition writing led to greater vocabulary learning from sentence writing and argued that strong evaluation needs to be revised, the findings of Kim (2008)

indicated that both activities led to similar vocabulary learning gains and suggested that such revision might not be needed. The present thesis took advantage of meta-analysis and the information theoretic approach to identify the most useful predictor variables, updated the ILH by adding other variables that contribute to the prediction, and revised the components of the ILH.

The ILH provided a systematic framework with which one could quantify multiple factors related to activity features. The current thesis revealed the relative effects of factors that contributed to incidental vocabulary learning. It was found that evaluation was the most influential factors among the components of the ILH and need contributed to learning to a lesser extent. In contrast, search did not clearly promote learning. These findings are not only pedagogically valuable but also theoretically meaningful because they enable researchers to discuss multiple factors simultaneously. Such discussion helps to elaborate upon a complex framework that helps to explain L2 incidental vocabulary learning.

5.2.2 Pedagogical Implications

The findings of this thesis produced several pedagogical implications. First, it is essential for language teachers to help students efficiently acquire vocabulary by ensuring that they engage in effective activities (Nation, 2007). The present thesis's findings indicate that productive activities, where students use target words in an original sentence or composition, are effective activities for word learning. Study 3 revealed that composition-level varied use activities are likely to lead to the greatest vocabulary learning of the activity types examined. Therefore, activities such as composition-writing using a set of target words, writing a letter, writing a speech transcript, and retelling activities are recommended to be included in the classroom or as homework.

Second, one way to enhance vocabulary learning is to design activities to include factors that have been determined to contribute to vocabulary learning (Coxhead, 2018; Nakata & Webb, 2017; Webb & Nation, 2017). Therefore, creating situations that include activity types that increase learning should be encouraged. For example, combining MFI (reading or listening) with writing activities may enhance vocabulary retention compared

to when learners only encounter target words through MFI. Similarly, it may also be useful to increase the frequency of encounters or use of target words by repeating the same activity or similar activities (e.g., Folse, 2006). When spoken activities (listening or speaking) are implemented, combining them with written activities (reading or writing) may also be advised (e.g., Brown et al., 2008; Jin & Webb, under review).

Third, although MFI was found to be the least effective activity type in terms of vocabulary learning, this finding does not mean that MFI activities should be abandoned. The findings of this thesis instead provided an important message that teachers and learners should not expect great vocabulary learning to occur through MFI over a short period of time. Study 3 revealed that there is only a small chance that learners would recall more than 50% of unknown words even when they are tested immediately after the activity. Research has demonstrated that knowledge of vocabulary is gained in small increments through repeatedly encountering words (Webb & Cheung, 2015; see also Webb, 2020). Therefore, it may be important to have reasonable expectations of learning so that learners can plan and continue learning without getting disappointed by a lack of immediate learning gains.

Lastly, a useful finding from Study 1 was that when predicting the effectiveness of activities, the IL of activities may be a more useful variable to consider compared to time on task. Although activities taking longer tended to lead to greater learning, this advantage is likely to be an artifact of increased ILs. If the focus is to foster vocabulary knowledge, teachers and learners should be encouraged to select activities with higher ILs rather than choosing activities that take longer.

5.3 Future Directions

5.3.1 Areas that Require Attention to Further Investigate the ILH

Through meta-analyzing earlier studies that tested the ILH, this thesis identified several areas that requires further attention in research. First, none of the ILH studies examined a learning condition involving strong need, where learners choose unknown words to pursue the goals of their tasks. Although the effects of motivation on vocabulary

learning have been discussed (e.g., Tseng & Schmitt, 2008), few studies examined how different manipulations of motivation-related factors influence the effect of tasks. Future studies should examine how motivational factors impact vocabulary learning through manipulating the need component.

Second, it was found that the majority of studies focused on single word learning (however, see Cao, 2013; Snoder, 2017), thus it remains unclear whether this thesis's evaluation of the predictive ability of the ILH can be generalized to multiword item learning. Similarly, more studies are required to examine the relationship between vocabulary knowledge and the ILH by measuring different aspects of vocabulary knowledge. Learning gains were mainly measured with form-meaning connection tests (e.g., multiple-choice tests or translation tests). Although other test formats have been administered, it is still unclear how other aspects of vocabulary knowledge (e.g., associations, collocations, pronunciations , spellings, and constraints on use) develop through engaging in tasks (e.g., Webb, 2005; see also Yanagisawa and Webb, 2020, for a review of different approaches to measuring depth of vocabulary knowledge).

Lastly, most studies had either no repetitions with target items or a small number of repetitions, making it difficult to clarify how frequency interacts with the effect of conditions present in tasks. Studies strictly examining the effect of frequency while manipulating the IL of tasks are still scarce (e.g., Eckerth and Tavakoli, 2012). Further research is needed to investigate whether the effect of IL changes as frequency increases.

5.3.2 Limitations Related to the Present Thesis and Future Directions

Although the findings of the three studies in this thesis are useful, the current thesis also suffers from several limitations.

5.3.2.1 Limitations Related to the Updated ILH

First, the updated ILH and IL formulas proposed in the present thesis are based on a simple predictive model. The IL formulas did not include the effects of interactions between variables. The effect of a particular variable might change based on other

variables. For instance, the effects of composition-level varied use might be more strongly pronounced when learning gains are tested with productive tests (e.g., form recognition) compared to receptive tests (e.g., meaning recognition). Additionally, the effect of some factors could decrease or increase when combined with other factors. For instance, the effect of frequency could be more strongly observed when an activity involves composition-level varied use compared to when an activity involves evaluation (Uchihara et al., 2019). Future research needs to examine different combinations of factors in order to deepen our understanding of how factors interact with each other to impact vocabulary learning.

Furthermore, it should be noted that a limited number of predictor variables (e.g., ILH components, frequency, mode, test format, test day) were examined through creating the updated ILH. Other factors potentially contributing to the prediction of learning gains (e.g., L2 proficiency, Kim, 2008; target word characteristics, Ellis & Beaton, 1993) were not included in the analysis because the current thesis adopted the theoretic-information approach, with which all predictor variables need to be reported in all included studies. To fully take advantage of individual studies, it would be useful for future studies to make their datasets and materials (e.g., target words and reading texts) publicly available. Open materials and datasets will help future meta-analyses code a greater number of predictor variables and examine them more accurately. Furthermore, future research should investigate other factors that are not included in the updated ILH. Examples of such factors include the characteristics of learners (e.g., proficiency, Kim, 2008; working memory, Yang, Shintani, & Zhang, 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).

Lastly, it is important to keep in mind that the updated ILH and IL formulas are representative of the studies that were analyzed. These studies, however, represent a limited set of possible combinations of predictor variables and the components of the ILH. For instance, some variables of the IL formula (i.e., frequency, mode, and search) were not all comprehensively examined with different conditions involving varying

levels of the ILH components (i.e., need, evaluation, sentence-level varied use, composition-level varied use). Therefore, future research needs to examine the accuracy of predictions of the updated ILH by investigating a variety of learning conditions with a greater combination of factors.

5.3.2.2 Limitations Related to Predicting Incidental Learning Gains

The present thesis provided a first attempt to propose a predictive model that estimates the amount of vocabulary learning gains based on learning conditions. The results of Study 2 and Study 3 provided a useful order of the efficacy of learning conditions, as well as the estimations of learning gains that will occur through completing different activity types. Given that the calculated predictive intervals of the estimated mean learning gains were relatively wide, it may be useful for future research to try improving the prediction by adding other predictor variables such as learner characteristics (e.g., prior vocabulary knowledge, Webb and Chang, 2015) and target word features (e.g., number of letters, pronounceability, imageability, concreteness, Ellis and Beaton, 1993).

5.3.2.3 Limitations Related to Incidental Vocabulary Learning

The final limitation relates to the fact that the thesis research exclusively focused on incidental vocabulary learning. Although this was a necessary first step, it may be useful to examine the extent to which the ILH can be generalized to deliberate vocabulary learning activities. For instance, while this thesis found no clear benefit of including search in learning, search might facilitate learning in a context of deliberate learning. One potential reason for this is that when information about target words is not at learners' disposal and they have to search e.g., by using a dictionary, they may try to recall it from memory. Research has demonstrated that such retrieval attempts tend to enhance vocabulary learning (Barcroft, 2007; Nation & Webb, 2011; Rott, 2007). Potential applications of the ILH to deliberate learning have been discussed (Nation & Webb, 2011), but never systematically investigated by examining learning gains reported in

empirical studies. Therefore, it would be useful for future meta-analyses to examine the predictive power of the ILH on vocabulary learning in the realm of deliberate learning.

5.4 Conclusion

This dissertation investigated the ILH through meta-analyzing studies that tested the ILH. Study 1 examined the predictive power of the ILH, the relative effects of the different components of the ILH, and the interaction between IL and other empirically motivated variables (e.g., frequency and time on task). Study 2 aimed to update the ILH to enhance its accuracy in predicting the relative effects of incidental vocabulary learning activities. Based on the results, we created an Involvement Load (IL) formula and proposed an updated ILH, with which one can calculate the updated ILs of activities to more accurately predict their relative effectiveness on incidental vocabulary learning. Study 3 categorized different incidental vocabulary learning conditions that have been examined in studies of the ILH and obtain the estimated vocabulary learning gains that occur across those activities.

Taken together, the contribution of the present thesis can be summarized as follows:

1. Deepened understanding of the relative effects and interactions of different factors contributing to incidental vocabulary learning
2. Enhanced predictions of the relative effectiveness of activities
3. Revealed how accumulated research findings could be more easily applied to pedagogy by estimating learning gains for different incidental vocabulary learning activities.
4. Identified topics requiring further attention in studies of L2 incidental vocabulary learning.

It is hoped that more research will be conducted to evaluate the accuracy of the predictions of the updated ILH and further enhance predictions of incidental vocabulary learning.

5.5 References

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

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*(2), 136.

Cao, Z. (2013). The effects of tasks on the learning of lexical bundles by Chinese EFL learners. *Theory and Practice in Language Studies*, *3*(6), 957–962. https://doi.org/10.4304/tpls.3.6.957-962

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

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*(2), 227–252. https://doi.org/10.1177/1362168811431377

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic Determinants of Foreign Language Vocabulary Learning. *Language Learning*, *43*(4), 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, *12*(3), 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*(2), 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 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*(1), 89–108. https://doi.org/10.1177/003368829702800106

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.

Nakata, T., & Webb, S. (2017). Vocabulary learning exercises: Evaluating a selection of exercises commonly featured in language learning materials. In B. Tomlinson, University of Liverpool, & Materials Development Association (United Kingdom) (Eds.), *SLA research and materials development for language learning*. Routledge.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt039.0

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

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, *57*(2), 165–199. https://doi.org/10.1111/j.1467-9922.2007.00406.x

Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of involvement load, spacing, and intentionality. *TESL Canada Journal*, *34*(3), 140–164. https://doi.org/10.18806/tesl.v34i3.1277

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*(2), 357–400. https://doi.org/10.1111/j.1467-9922.2008.00444.x

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

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2020). Incidental Vocabulary Learning. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 225–239). Routledge.

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

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

Yanagisawa, A., & Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 371–386). New York, NY: Routledge. https://doi.org/10.4324/9780429291586-24

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

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*(1), 54–75. https://doi.org/10.1177/1362168816652418

Appendices

Appendix A: Basic Information about Included Studies

study	Target language	Activity	Need	Search	Evaluation	ILH	Test format
Ansarin&Bayazidi2016	English	Multiple choice	1	1	1	3	Receptive (meaning) recall
		Gap-filling	1	1	1	3	
		Writing	1	1	2	4	
Baleghizadeh&Abbasi2013	English	Reading	1	0	0	1	Receptive (meaning) recall & Productive use
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
		Writing	1	0	2	3	

Bao2015	English	Reading	0	0	0	0	Receptive (meaning) recall & Productive use
		Matching	1	0	1	2	
		Sentence combination	1	0	1	2	
		Translation	1	0	1	2	
		Writing	1	0	2	3	
Beal2007	English	Reading	0	0	0	0	Receptive (meaning) recall & VKS
		Reading	1	0	0	1	
		Reading	1	0	1	2	
		Writing	1	1	2	4	
Cao2013	English	Reading	1	0	0	1	Receptive (meaning) recall
		Gap-filling	1	0	1	2	

		Writing	1	0	2	3	
Cheng2011	English	Reading	1	0	0	1	VKS
		Gap-filling	1	1	1	3	
		Writing	1	1	2	4	
Chenghai&Feng2017	English	Reading	1	0	0	1	Receptive (meaning) recall
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
Folse2006	English	Gap-filling	1	1	1	3	VKS
		Gap-filling	1	1	1	3	
		Writing	1	1	2	4	
Hazrat2015	English	Reading	1	0	1	2	Productive (form) Recall & Receptive (meaning) recall
		Writing	1	0	2	3	

		Speaking	1	0	2	3	
Hirata&Mori2008	English	Multiple choice	1	0	1	2	Receptive (meaning) recall & Productive (form) Recall
		Gap-filling	1	0	1	2	
Hulstijn&Laufer2001	English	Reading	1	0	0	1	Receptive (meaning) recall
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
Hyun2011	English	Reading	1	1	1	3	Receptive (meaning) recall
		Gap-filling	1	1	1	3	
		Writing	1	0	2	3	
Jahangard2013	English	Listening	1	0	0	1	Productive use
		Listening	1	1	0	2	
		Writing	1	1	2	4	

Jahangiri&Abilipour2014	English	Writing	1	1	2	4	VKS
		Gap-filling	1	1	1	3	
		Writing	1	1	2	4	
		Gap-filling	1	1	1	3	
Jing&Jianbin2009	English	Listening	0	0	0	0	Receptive (meaning) recall
		Listening	1	0	0	1	
		Writing	1	0	2	3	
Karalik&Merç2016	English	Gap-filling	1	0	1	2	Receptive (meaning) recall
		Gap-filling	1	1	1	3	
		Speaking	1	0	2	3	
		Speaking	1	1	2	4	
Keyvanfar&Badraghi2011	English	Reading	1	0	0	1	Receptive (meaning) recall
		Gap-filling	1	0	1	2	

		Writing	1	0	2	3	
Khonamri&Hamzenia2013	English	Writing	1	0	2	3	Receptive (meaning) recall
		Gap-filling	1	1	1	3	
		Translation	1	1	1	3	
Kim2008	English	Reading	1	0	0	1	VKS
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
		Writing	1	0	2	3	
Kolaiti&Raikou2017	English	Reading	1	1	0	2	Receptive (meaning) recall
		Reading	1	0	0	1	
Konno et al.2009	English	Gap-filling	1	0	1	2	VKS
		Writing	1	0	2	3	
		Gap-filling	1	1	1	3	

		Writing	1	1	2	4	
Lee&Hirsh2012	English	Multiple choice	1	0	1	2	VKS
		Multiple choice	1	0	1	2	
		Writing	1	0	2	3	
Li2014	English	Reading	0	0	0	0	Receptive (meaning) recall
		Reading	1	1	0	2	
		Gap-filling	1	1	1	3	
		Writing	1	1	2	4	
Maleki2012	English	Listening	0	0	0	0	Receptive (meaning) recognition
		Listening	1	0	0	1	
		Writing	1	0	2	3	
Martínez-Fernández2008	Spanish	Reading	1	0	0	1	Productive (form) Recall & Receptive (meaning) recall

							& Form recognition & Receptive (meaning) recognition
		Gap-filling	1	0	1	2	
		Reading	1	0	1	2	
Rott2012	German	Reading	1	0	0	1	Receptive (meaning) recall & Productive (form) Recall
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
Sarbazi2014	English	Reading	0	0	0	0	Receptive (meaning) recall
		Reading	1	0	0	1	
		Writing	1	0	2	3	
Snoder2017	English	Gap-filling	1	0	1	2	Productive (form) Recall
		Writing	1	0	2	3	
		Gap-filling	1	1	1	3	

		Writing	1	1	2	4	
Soleimani&Rahmanian2014	English	Gap-filling	1	0	1	2	Receptive (meaning) recall
		Writing	1	0	2	3	
Soleimani&Rahmanian2015	English	Gap-filling	1	0	1	2	Receptive (meaning) recall
		Reading	1	0	0	1	
Tang&Treffers-Daller2016	English	Reading	0	0	0	0	Receptive (meaning) recall
		Reading	0	0	0	0	
		Reading	1	1	0	2	
		Reading	1	0	1	2	
		Reading	1	1	1	3	
		Writing	1	0	2	3	
Teng2015b	English	Reading	1	0	0	1	VKS
		Gap-filling	1	0	1	2	

		Writing	1	0	2	3	
		Writing	1	1	2	4	
Teng2015c	English	Reading	1	0	0	1	Receptive (meaning) recall & Productive use
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
Teng2017a	English	Reading	1	1	0	2	VKS
		Gap-filling	1	1	1	3	
		Writing	1	1	2	4	
Teng2017b	English	Reading	1	0	0	1	Receptive (meaning) recall
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	

Tsubaki2012	English	Graphic organizers	1	0	2	3	Productive (form) Recall & Receptive (meaning) recognition
		Graphic organizers	1	0	1	2	
Tu2004	English	Reading	1	0	0	1	Receptive (meaning) recall
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
Wang et al.2014	English	Reading	1	1	0	2	Receptive (meaning) recall
		Matching	1	0	1	2	
		Writing	1	0	2	3	
Yang et al.2017	English	Writing	1	0	2	3	VKS
		Gap-filling	1	0	1	2	
		Reading	0	0	0	0	

Yang2015	English	Reading	1	0	0	1	Receptive (meaning) recall & Productive (form) Recall
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
Yaqubi et al.2010	English	Reading	1	1	1	3	Receptive (meaning) recall
		Gap-filling	1	0	1	2	
		Writing	1	0	2	3	
Zou2017	English	Gap-filling	1	0	1	2	VKS
		Writing	1	0	2	3	
		Writing	1	0	2	3	

Appendix B: Coding Scheme for Study 1

Coding column	Explanations of the column	Possible responses	Notes
study_no			
author			
year			
study			
exp			
participant_group			
publication_type		(1) journal - research journals (2) PhDthesis (3) MAthesis (4) bulletin - university journals (5) conference - conference	

	presentation; conference preceeding	
region		
L1		
target_language		
institution	(1) elementary (2) secondary (3) university (4) language_school	Pre-university students in a certain language program were coded as language_school (Folse, 2006; Kim, 2008) When the research was carried out in a language institution even their institutional level is high school, it was coded as

			language institution. (i.e., Jahangiri & Abilipour2014)
activity	Type of activity		
activity2	Larger category of the type of activity	(1) fill_in (2) translation (3) writing (4) reading (5) graphic_organizers (6) matching (7) multiple_choice (8) speaking	(6) matching -any forms of matching activity (follow the authors' labeling: it does not matter what cognitive processes involved: coded just based on the format of an activity where participants were asked to match two items). (7) multiple_choice - any forms of multiple-choice activities (follow the authors' labeling: it

			does not matter what cognitive processes involved: coded just based on the format of an activity where participants were provided with multiple-choices and asked to select the most appropriate one). (8) speaking - oral-sentence-production (Hazrat, 2015) and retelling (Karalik & Merç, 2016)
author.need	a need component reported by the author	0 - no need 1 - moderate need 2 - strong need	
author.search	a search component reported by the author	0 - no search 1 - search was present	

author.evaluation	an evaluation component reported by the author	0 - no evaluation 1 - moderate evaluation 2 - strong evaluation
author.ILH	Total task-induced involvement load index reported by the author	
does_authors_coding_followed_ILH_exactly		1 - Yes 0 - No
need	a need component re-coded by the meta-analysts	0 - no need 1 - moderate need 2 - strong need
search	a search component re-coded by the meta-analyst	0 - no search 1 - search was present
evaluation	a evaluation component re-coded by the meta-analyst	0 - no evaluation 1 - moderate evaluation 2 - strong evaluation

ILH	Total task-induced involvement load index re-coded by the meta-analyst	
activity_time	Minutes (mean or median) participants engaged in the learning condition	
time_per_word	Activity time (Minutes) divided by the number of target words	
frequency	The number of times participants encountered or used the same set of target words	
posttest_announcement	Whether participants were told that they will be tested for vocabulary after the treatment	Coded as "NR" when the authors did not specify

number_of_target_words	Number of target words participants were exposed in the treatment (i.e., learning condition) in question	
how_to_make_sure_learners_did_not_know_target_words		(a) Pretest (b) Pilot study and/or other students (c) Consulting with teachers (d) Considering word frequency (e) After treatment questionnaire (f) test-only group (g) Nonword use NR - Not reported

pre_test_administration	Whether pretests were administered to measure participants' prior knowledge of target words	0 - no 1 - yes	
prior_knowledge_control	Whether participants' prior knowledge of target words were directly controlled by one or more than one of the method in the 'how to make sure learners did not know target words' column	1 - prior knowledge was controlled by using any ways 0 - it was not clear how the authors made sure/controlled participants' prior knowledge of the target words: 'how to make sure learners did not know target words' column was 'NR'	Coded as 0 when the column 'how to make sure learners did not know target words' was coded as NR Coded as 1 when the column 'how to make sure learners did not know target words' was coded as anything else instead of NR

vocabulary_item_type	Type of to-be-learned items	(1) single words (2) multiword units (3) mix
reliability_reported	Whether statistical reliability scores (e.g., Cronbach's alpha) of the posttest scores were reported	
test_format		(1) receptive recall - e.g., translation (L2-> L1) (2) receptive recognition (3) productive recall - e.g., translation (L1-> L2) (4) productive recognition (5) form recognition - recognize whether target words were present in the treatment

	(6) VKS (7) gap-filling (8) productive use - when participants were asked to write a sentence using a target word and the sentence was judged based on its semantic and grammatical accuracy; or just asked to use (Feng, 2015)
test_format2	(1) recall (2) recognition - e.g., multiple-choice tests (3) other - VKS, gap-filling, productive-use

tests_max_score	Maximum score for the test	
pretest_mean		For VKS, 1 point x the number of target words was inserted when VKS's Category I scored 1 point.
pretest_test_SD		
how_many_days_until_the_immediate_test	Number of days between treatment and the immediate posttest	
immediate_test_n	Number of participants who took the test in question	
immediate_test_M		
immediate_test_SD		

how_many_days_until_the_delayed_test	Number of days between treatment and the delayed posttest
delayed_test_n	Number of participants who took the test in question
delayed_test_M	
delayed_test_SD	
immediate_test_ES	ES = $(posttest\ score - pretest\ score)/(test\ score\ maximum - pretest\ score)$
immediate_test_ES_SD	ES = (posttest score SD)/(test score maximum)
delayed_test_ES	ES = $(posttest\ score - pretest\ score)/(test\ score\ maximum - pretest\ score)$

delayed_test_ES_SD	ES = (posttest score SD)/(test score maximum)

Appendix C: Coding Scheme for the ILH Components

Criteria	Definitions (Laufer & Hulstijn, 2001)	Examples of learning conditions (Laufer & Hulstijn, 2001)	Examples in other related-articles	Tricky cases: Coding decision
Need	"The need component is the motivational, non-cognitive dimension of involvement. It is concerned with the *need to achieve.* We interpret this notion not in its negative sense, but in its positive sense, based on a drive to comply with the task requirements, whereby the task requirements can be either externally imposed or self-imposed" (p. 14)	"If, for example, the learner is reading a text and an unknown word is absolutely necessary for comprehension, s/he will experience the need to understand it" (p. 14) "Or, the need will arise during a writing or speaking task when the L2 learner wants to refer to a certain concept or object but the L2 word expressing it is unfamiliar" (p. 14) "A reading comprehension task which requires the learner to look up the meaning of a homonym in a dictionary illustrates need (since		

	knowing the word's meaning is necessary for the successful completion of the comprehension task), search (since the meaning of the word is looked up), and evaluation (since different meanings of the word have to be compared and checked against the context before one is selected)" (p. 15)
Need: absent	"A reading comprehension task where unknown words are glossed for the student, but the comprehension questions can be answered without reference to these words does not induce any need to focus on the glossed words (since they are irrelevant to the task), nor any search for their meaning (since

		they are glossed), nor any evaluation" (p. 16)	
Need: moderate	"Need is moderate when it is imposed by an external agent, e. g. the need to use a word in a sentence which the teacher has asked the learner to produce" (p 14)	"If, [] the same task [a reading comprehension task which requires the learner to look up the meaning of a homonym in a dictionary] is simplified for the learner by teacher's glosses for unknown words in the text margin, search and evaluation are no longer required. In [this] example, the task induces a weaker involvement as only the need component is at work" (p. 15) "A reading comprehension task with glossed words that are relevant to answering the questions will induce a moderate need to look at the glosses (moderate because it is	For reading tasks, it was difficult to determine whether or not participants needed to understand target words to complete the task when the study did not clearly explain. For example, some studies did not clearly state that in reading conditions, target words had to be understood in order to comprehend the text or the participants had

	imposed by the task), but it will induce neither search nor evaluation" (p. 15) "The same task [a reading comprehension task] with glosses removed [assuming that a dictionary is provided] will not only induce need but also search (provided that the student has deemed the unknown words as relevant enough to look up)" (p. 15) "The fill-in [the blanks in a text] task induces a moderate need, no search (the words are explained) and a moderate evaluation, since all the words in the list have to be evaluated against each other and the context of the gaps" (p. 17)	to answer comprehension questions that required them to understand the target words. In such cases, we contacted authors of the studies for clarification. When detailed information was not provided by the authors, need was coded as the coding found in each study. However, when studies coded need as moderate, but how the participants understand the meanings of target

		words (e.g., guessing while reading) was not clearly stated, we contacted the authors for clarification. When detailed information was not provided by the authors, the study was excluded.
	"[] the learner is asked to write original sentences with some new words. These words are translated or explained by the teacher. The task induces a moderate need, no search, and strong evaluation because the new words are evaluated against suitable collocations in a learner-generated context" (p. 17)	
	"[] the learner is required to write a composition [] and incorporate some L2 target words; the teacher has not provided these words in their L2 form, but by their L1 equivalent [and the learner use a dictionary to look up L2 word forms]. The task will induce a moderate need and search since the L2 word forms have	In order to examine the influence of studies that coded need as moderate for their reading condition but did not explain why the participants had to understand target words, we ran sensitivity analyses

	to be looked up, and again a strong evaluation as the words are used in learner-generated context" (p. 17)	and confirmed the influence may be none to negligible (see Appendix E).

Need: strong	"Need is strong when imposed on the learner by him- or herself" (p. 14)	"Consider a case of a composition where the learner wants to use concepts for which s/he possesses no L2 form. S/he then decides to look up these L1 concepts for their L2 equivalence (in an L1-L2 dictionary) and use them in the composition. This task induces a strong need (self-imposed), search, and a strong evaluation" (p. 17) "The input task is to read a text for comprehension. During the reading, the learner decides to look up certain words in a dictionary. Since it was the learner's decision, the need is characterised as strong" (p. 20)	Need was coded as strong only when the desire to use a certain word is purely motivated by participants. Therefore, when researchers made a learning condition require participants to understand/use a certain set of target words (e.g., Wang et al., 2014, where target words were essential to understand the passage and to answer true/false questions), need was coded as moderate.

Search	"Search is the attempt to find the meaning of an unknown L2 word or trying to find the L2 word form expressing a concept (e.g. trying to find the L2 translation of an L2 word) by consulting a dictionary or another authority (e.g. a teacher)" (p. 14)	"A reading comprehension task which requires the learner to look up the meaning of a homonym in a dictionary illustrates need (since knowing the word's meaning is necessary for the successful completion of the comprehension task), search (since the meaning of the word is looked up), and evaluation (since different meanings of the word have to be compared and checked against the context before one is selected)" (p. 15)

Search: absent	"If, [], the same task [a reading comprehension task which requires the learner to look up the meaning of a homonym in a dictionary] is simplified for the learner by teacher's glosses for unknown words in the text margin, search and evaluation are no longer required. In [this] example, the task induces a weaker involvement as only the need component is at work" (p. 15) "A reading comprehension task where unknown words are glossed for the student, but the comprehension questions can be answered without reference to these words does not induce any need to focus on the glossed words (since they are irrelevant to the task), nor	Reading with multiple-choice glosses conditions were coded as not involving search. This is based on the description in Laufer and Hulstijn (2001, p. 18-19): "Hulstijn (1992) showed that when meanings of words had to be inferred they were retained better than words with given meanings. If we compare the two tasks in terms of involvement load, we can see that the

	any search for their meaning (since they are glossed), nor any evaluation" (p. 16)	difference lies in the absence of evaluation in the synonym-condition and presence of evaluation in the multiple-choice condition. Learners had to evaluate all the alternative meanings against the text context. (In both conditions there was a moderate need, induced by the researcher, and no search)."
	"A reading comprehension task with glossed words that are relevant to answering the questions will induce a moderate need to look at the glosses (moderate because it is imposed by the task), but it will induce neither search nor evaluation" (p. 15)	
	"When unknown words are not negotiated, it means the learner has no need for them and therefore performs no search" (p. 19)	
	"The fill-in [the blanks in a text] task induces a moderate need, no	

	search (the words are explained) and a moderate evaluation, since all the words in the list have to be evaluated against each other and the context of the gaps" (p. 17)
	"[] the learner is asked to write original sentences with some new words. These words are translated or explained by the teacher. The task induces a moderate need, no search, and strong evaluation because the new words are evaluated against suitable collocations in a learner-generated context" (p. 17)

Search: moderate	"The same task [a reading comprehension task] with glosses removed [assuming that a dictionary is provided] will not only induce need but also search (provided that the student has deemed the unknown words as relevant enough to look up)" (p. 15) "[] the learner is required to write a composition [] and incorporate some L2 target words; the teacher has not provided these words in their L2 form, but by their L1 equivalent [and the learner use a dictionary to look up L2 word forms]. The task will induce a moderate need and search since the L2 word forms have to be looked up, and again a strong evaluation as the words are used in	Search might not be dichotomously coded as either present or absent because most of the learning conditions include a search component to some extent when participants were provided with some kind of material that presents information about target words. We coded learning conditions for search following each study's operationalization of search. That is, search was coded as the

	learner-generated context" (p. 17) "Consider a case of a composition where the learner wants to use concepts for which s/he possesses no L2 form. S/he then decides to look up these L1 concepts for their L2 equivalence (in an L1-L2 dictionary) and use them in the composition. This task induces a strong need (self-imposed), search, and a strong evaluation" (p. 17) "Search for meaning does not have to be in a dictionary only. The learner can search the text context, ask a teacher, or peers" (p. 19)	coding found in each study unless its operationalization did not follow Laufer and Hulstijn's (2001) description of the ILH. As a result, the following conditions were coded as including search: when a mini dictionary was provided (e.g., Folse, 2006), when hyperlinked glosses were provided (Li, 2014), when Google translator was used (e.g., Kolaiti & Raikou, 2017), when a

		glossary was provided at the end of the text and meanings of the words were ordered alphabetically (i.e., Tang & Treffers-Daller, 2016). No other methods such as searching the meaning by examining the context around the word or consulting with a teacher or peers were operationalized as search.

Evaluation	"Evaluation entails a comparison of a given word with other words, a specific meaning of a word with its other meanings, or combining the word with other words in order to assess whether a word (i.e. a form-meaning pair) does or does not fit its context" (p. 14) "Evaluation [] implies some kind of selective decision based on a criterion of semantic and formal appropriateness (fit) of the word and its context" (p. 15)	"If, for example, during a reading task, a word that is looked up is a homonym, a decision has to be made about its meaning by comparing all its meanings against the specific context and choosing the one that its best" (p. 14) "Another example is an L2 writing task in which an L1 word is looked up in a dictionary and three L2 alternatives are presented. The translations have to be evaluated against each other and the most suitable one has to be chosen for the specific meaning the L2 writer is trying to convey. But unlike in the preceding example, the evaluation in the writing task will involve additional syntagmatic decisions

about the precise collocations of the word which the learner is trying to use" (p. 15)

Evaluation: absent	"[During reading], If unknown words have only one meaning and if the context allows a straight forward, literal interpretation of it, no decision has to be made about its contextual meaning" (p. 16) "If, [], the same task [a reading comprehension task which requires the learner to look up the meaning of a homonym in a dictionary] is simplified for the learner by teacher's glosses for unknown words in the text margin, search and evaluation are no longer required. In [this] example, the task induces a weaker involvement as only the need component is at work" (p. 15) "A reading comprehension task

where unknown words are glossed for the student, but the comprehension questions can be answered without reference to these words does not induce any need to focus on the glossed words (since they are irrelevant to the task), nor any search for their meaning (since they are glossed), nor any evaluation" (p. 16)

"A reading comprehension task with glossed words that are relevant to answering the questions will induce a moderate need to look at the glosses (moderate because it is imposed by the task), but it will induce neither search nor evaluation" (p. 15)

	"The same task [a reading comprehension task] with glosses removed [assuming that a dictionary is provided] will not only induce need but also search (provided that the student has deemed the unknown words as relevant enough to look up)" (p. 15)

Evaluation: moderate	"If the evaluation entails recognising differences between words (as in a fill-in task with words provided), or differences between several senses of a word in a given context, we will refer to this kind of evaluation as 'moderate'" (p. 15)	"If, on the other hand, the word has several meanings, the reader has to select the meaning which makes sense in the context, a decision demanding moderate evaluation -- moderate since the learner is not required to produce original language" (p. 16) "The fill-in [the blanks in a text] task induces a moderate need, no search (the words are explained) and a moderate evaluation, since all the words in the list have to be evaluated against each other and the context of the gaps" (p. 17) "[While reading a text with multiple-choice glosses], meaning selected from several options. []	"[] a sentence can be translated in more than one way. The final choice of the translation must have been made after an evaluation of several translation alternatives. In each option, the target word was evaluated against the other words surrounding it. Moreover, in L1-L2 translation, the entire L2 context was created by the learner. Hence, the element of **'evaluation' was moderate in the L2-L1 translation task** and strong in the L1-L2	For reading with a dictionary, when the target words were polysemous, there would be a moderate evaluation component. However, it may depend on the type of target words as well as the type of dictionaries participants used. Therefore, we followed each study's coding of evaluation (e.g., Yaqubi et al., 2010: moderate evaluation; Wang et al., 2014: no evaluation)

	learners had to evaluate all the alternative meanings against the text context. (in both conditions there was a moderate need, induced by the researcher, and no search)" (p. 19).	translation" (Laufer & Girsai, 2008, p. 712)

Evaluation: strong	"If, on the other hand, evaluation requires making a decision about additional words which will combine with the new word in an original sentence or text, we will refer to it as 'strong' evaluation" (p. 15)	"[] the learner is asked to write original sentences with some new words. These words are translated or explained by the teacher. The task induces a moderate need, no search, and strong evaluation because the new words are evaluated against suitable collocations in a learner-generated context" (p. 17) "[] the learner is required to write a composition [] and incorporate some L2 target words; the teacher has not provided these words in their L2 form, but by their L1 equivalent [and the learner use a dictionary to look up L2 word forms]. The task will induce a moderate need and search since the L2 word forms have to be looked up, and again a strong	"[] a sentence can be translated in more than one way. The final choice of the translation must have been made after an evaluation of several translation alternatives> In each option, the target word was evaluated against the other words surrounding it. Moreover, in L1-L2 translation, the entire L2 context was created by the learner. Hence, the element of **'evaluation' was** moderate in the L2-L1 translation task and **strong in the L1-L2**

	evaluation as the words are used in learner-generated context" (p. 17) "Consider a case of a composition where the learner wants to use concepts for which s/he possesses no L2 form. S/he then decides to look up these L1 concepts for their L2 equivalence (in an L1-L2 dictionary) and use them in the composition. This task induces a strong need (self-imposed), search, and a strong evaluation" (p. 17)	**translation**" (Laufer & Girsai, 2008, p. 712)

Appendix D: Calculation Formulas for ESs and *SDs*

(a) Studies in which participants were exposed to target words that were all unknown to them during the treatment:

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

$$SD = \frac{SD\ of\ posttest\ score}{Maximum\ posttest\ score}$$

(b) For studies administering a pretest to measure participants' prior knowledge of the target words:

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

$$SD = \frac{SD\ of\ posttest\ score}{Maximum\ posttest\ score - Mean\ preposttest\ score}$$

(c) For studies that did not administer pretests but included a control group (i.e., a group that only took posttests without going through a treatment):

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

$$SD = \frac{SD\ of\ posttest\ score}{Maximum\ posttest\ score - Mean\ preposttest\ score}$$

(d) For studies that used Vocabulary Knowledge Scale with which Category I scored 1 point and participants were exposed to target words that were all unknown to them during the treatment:

$$\text{ES} = \frac{Mean\ posttest\ score - Number\ of\ target\ words \times 1}{Maximum\ posttest\ score - Number\ of\ target\ words \times 1}$$

$$SD = \frac{SD\ of\ posttest\ score}{Maximum\ posttest\ score - Number\ of\ target\ words \times 1}$$

(e) For studies that used Vocabulary Knowledge Scale with which Category I scored 1 point and administered a pretest to measure participants' prior knowledge of the target words:

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

$$SD = \frac{SD\ of\ posttest\ score}{Maximum\ posttest\ score - Mean\ preposttest\ score}$$

All formulas for ESs were devised based on the formula of proportion of unknown words learned used provided by Swanborn and de Glopper (1999) (see also, Card, 2012, p. 148). SDs on the posttest scores were divided by the proportion of unknown words to make the SDs on the same scale of relative learning gains. The converted SDs were used to calculate the sampling variance using the formula in Hox (2010, p. 209), s^2/n, where s refers to SD (see also Card, 2012, p. 150).

Appendix E: Sensitivity Analyses and Additional Analyses for Study 1

Publication Bias Analyses

Potential publication biases—studies reporting larger learning gains favoring the Involvement Load Hypothesis (ILH) (or contrasting with ILH) might have been easier to get published—could prevent us from accurately assessing the predictive ability of ILH.

To account for such a potential publication bias, we included both published and unpublished studies (masters' theses, doctoral dissertations, and conference presentations) in our analysis. The majority of the included studies, 34 (81.0%), were published studies (30 research journal articles, 3 book chapters, and 1 research bulletins), and 8 (19.0%) were unpublished studies (4 MA theses, 2 Ph.D. dissertations, and 2 conference papers).

In order to test whether the status of publication (published or unpublished) is related to learning gains or the effect of Involvement Load (IL), we followed Card (2012, p. 262) and used a meta-regression model including the main effect of Publication status and the main effect of IL and an interaction between Publication status and IL. The analyses of immediate and delayed posttests showed that there were no significant main effects (p = .529, p = .702, respectively) or interaction effects (p = .648, p = .562, respectively). This indicates that neither the percentage of learning gains nor the effect of ILH may have influenced whether a study was published or not.

Moreover, there is a possibility that studies including a larger number of participants could have been more likely to get published (e.g., Huang et al., 2012). In order to evaluate this potential bias, we conducted Egger's type meta-regression analysis (Egger et al., 1997). A meta-regression model including the main effect of the number of participants and the main effect of IL as well as an interaction between the two were administered with ESs on immediate posttests and delayed posttests, separately. The results of the immediate and delayed posttests showed that there were no significant main effects (p = .358, p = .239, respectively) or interaction effects (p = .391, p = .351,

respectively). These results indicate that there were none to negligible publication biases among the studies included in the current meta-analysis.

Sensitivity Analyses Regarding ILH Coding

In order to ensure the accuracy and consistency of ILH coding, we devised a clear and detailed coding scheme (see Appendix C for the coding scheme). All potential questions regarding coding were solved through personal discussion with Batia Laufer (2019, personal communication). Furthermore, we evaluated potential influences regarding ILH coding by conducting additional sensitivity analyses as follows.

Influence of authors' coding. The process of the current meta-analysis revealed that there is some inconsistency among many authors' IL coding of conditions across studies. Eleven studies (26.2%) coded their learning conditions differently from Laufer and Hulstijn's (2001) description of ILH (see the completed coding scheme that is publicly available online). Furthermore, to examine the influence of the authors' coding of ILH, meta-regression analyses with IL predicting the percentage of learning gains were carried out again using the authors' ILH coding schemes. The results indicated that there was a trend showing that the explained variance by ILH slightly decreased when using an author's coding of ILH compared to when using strict coding according Laufer and Hulstijn's (2001) description (on immediate posttest, 9.6% of the total variance and 23.8% at the within-study level, and on delayed posttests, 2.7% and 21.7%, respectively, in contrast to when using the coding strictly following Laufer and Hulstijn, 2001, on immediate posttests, 15.0% of the total variance and 29.1% at the within-study level and on delayed posttests, 5.1% and 26.5%, respectively). However, the results of the meta-regression analysis including the main effect of IL and whether or not each study's coding strictly followed ILH and the interaction between the two did not find any statistically significant interaction effects (b = -0.036, p = .568, on immediate posttests and b = -0.068, p = .278, on delayed posttests). In sum, these results indicate that the effect of IL was smaller when each study's coding of IL did not strictly follow Laufer and

Hulstijn's (2001) description, but such an influence was not strong enough to obtain a statistical significance.

Influence of the ambiguity regarding a need component. We found some ambiguity in the coding of ILH among some of the reading conditions. Eleven studies included reading conditions where *need* was coded as moderate but did not clearly state whether the participants had to understand the target words for the comprehension of a text or answer comprehension questions. We contacted the authors of the eight studies (72.7%) for which we found contact information. We received replies from two authors (25%: Baleghizadeh & Abbasi, 2013; Teng & Zhang, 2015) and both reported that the participants had to answer comprehension questions that required them to understanding the meanings of the target words.

Furthermore, we reran the whole analysis while excluding the reading conditions where it was not clear whether participants needed to understand target words. The results indicated a similar trend of the data (i.e., similar coefficients, confidence intervals, and *p*-values), suggesting the influence of including these studies may be negligible, confirming the robustness of the results.

Different Operationalizations of the Search Component

Among 18 studies that included tasks where search was present, 16 studies operationalized search as dictionary look up (13 used paper dictionaries and 3 used electronic dictionaries including translation applications), and two studies operationalized search as glosses (Li, 2014, using hyperlinked glosses; Tang & Tang & Treffers-Daller, 2016, where a glossary was provided at the end of the text and meanings of the words were ordered in an alphabetical manner).

In order to examine whether the influence of search varied based on how search was operationalized, we categorized studies with search into groups: paper dictionary, electronic dictionary, and glosses (hyperlinked glosses and glossaries). The whole dataset was analyzed with meta-regression models predicting ESs including the different search

variables (i.e., paper dictionary, electronic dictionary, glosses) as well as need and evaluation as covariate variables. The results show that although all of the coefficients of different search operationalizations were negative, none of them were statistically significant (see Table 1). This indicates that no positive influence of search was observed across different operationalizations.

Table 1

Different Operationalizations of Search

	Immediate Posttests				Delayed Posttests			
	k	*n*	*b* [CI]	*p*	*k*	*n*	*b* [CI]	*p*
Paper Dictionary	12	31	-.049 [-.217, .119]	.345	8	20	-.029 [-.286, .227]	.399
Electronic Dictionary	3	5	-.020 [-.089, .048]	.504	2	3	-.049 [-.166, .069]	.344
Glosses	2	5	-.006 [-.422, .410]	.919	2	8	-.065 [-.536, .405]	.506

Notes. *k* = number of studies. *n* = number of ESs.

Appendix F: The Number of ESs for each Combination of Components of the ILH and Example of Activities

Number of ESs for each Combination of Components of ILH and Examples of Activities

ILH					
Need	Search	Evaluation	*n*	%	Examples of activities
0	0	0	20	5.0%	Reading, Listening
1	0	0	76	19.1%	Reading, Listening
1	0	1	127	31.9%	Fill-in-the-blanks, Matching, Translation, Reading with multiple-choice glosses
1	0	2	103	25.9%	Writing, Retelling, Speaking
1	1	0	12	3.0%	Reading with a dictionary, Listening with a dictionary, Reading with a glossary
1	1	1	34	8.5%	Fill-in-the-blanks using a dictionary
1	1	2	26	6.5%	Writing with a dictionary, Retelling with a dictionary

Notes. ILH = combination of components of ILH. *n* = number of ESs. 0 = absence of the component, 1 = moderate, 2 = strong.

Appendix G: References of Included Studies

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*(2), 135–146. https://doi.org/10.2989/16073614.2016.1201774

Baleghizadeh, S., & Abbasi, M. (2013). The effect of four different types of involvement indices on vocabulary learning and retention of EFL learners. *The Journal of Teaching Language Skills, 5*(2), 1–26.

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

Cao, Z. (2013). The effects of tasks on the learning of lexical bundles by Chinese EFL learners. *Theory and Practice in Language Studies, 3*(6), 957–962. https://doi.org/10.4304/tpls.3.6.957-962

Cheng, H.-C. (2011). *Vocabulary acquisition in learning English as a second language: Examining the involvement load hypothesis and language anxiety with Taiwanese college students* (Unpublished doctoral dissertation, University of Northern Colorado). Retrieved from http://search.proquest.com/openview/5bdbbde9d102cdab90e4ba4301b6ee3c/1?pq-origsite=gscholar&cbl=18750&diss=y

Chenghai, Q., & Feng, T. (2017). Assessing the correlation between task-induced involvement load, word learning, and learners' regulatory ability. *Chinese Journal of Applied Linguistics, 40*(3), 261–280. https://doi.org/10.1515/cjal-2017-0015

Feng Teng, M. (2017). Investigating task-induced involvement load and vocabulary learning from the perspective of metacognition. *Social Sciences & Humanities, 25*(4), 1753–1764.

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Hazrat, M. (2015). The effects of task type and task involvement load on vocabulary learning. *Waikato Journal of Education*, *20*(2), 79–92. https://doi.org/10.15663/wje.v20i2.189

Hirata, Y., & Mori, C. (2008). A study of effective tasks based on task-induced involvement in incidental vocabulary acquisition. *International Journal of Curriculum Development and Practice*, *10*(1), 25–37. https://doi.org/10.18993/jcrdaen.10.1_25

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Hyun, P. J. (2011). *The role of task-induced involvement load in vocabulary acquisition of Korean college students* (Unpublished master's thesis, Ewha Womans University). Retrieved from http://dspace.ewha.ac.kr/handle/2015.oak/188558

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.

Jahangiri, K., & Abilipour, I. (2014). Effects of collaboration and exercise type on incidental vocabulary learning: Evidence against involvement load hypothesis. *Procedia - Social and Behavioral Sciences*, *98*, 704–712. https://doi.org/10.1016/j.sbspro.2014.03.471

Jing, L., & Jianbin, H. (2009). An empirical study of the involvement load hypothesis in incidental vocabulary acquisition in EFL listening. *Polyglossia*, *16*, 1–11.

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*(35), 77–92.

Keyvanfar, A., & Badraghi, A. H. (2011). Revisiting task-induced involvement load and vocabulary enhancement: Insights from the EFL setting of Iran. *Man & the Word/Žmogus Ir Žodis*, *13*(3), 56–66.

Khonamri, F., & Hamzenia, Z. (2013). The role of task-induced involvement in vocabulary learning. *English Review: Journal of English Education*, *1*(2), 1–11.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, *58*(2), 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*(3), 406–428. https://doi.org/10.22158/selt.v5n3p406

Konno, K., Takanami, S., Okuyama, Y., & Hirai, A. (2009). Examining the effects of involvement load on Japanese EFL learners' vocabulary retention. *JLTA Journal*, *12*, 46–64. https://doi.org/10.20622/jltaj.12.0_46

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). https://doi.org/10.3726/978-3-0351-0379-3

Li, J. (2014). Effect of task-induced online learning behavior on incidental vocabulary acquisition by Chinese learners—revisiting involvement load hypothesis. *Theory and Practice in Language Studies*, *4*(7), 1385–1394. https://doi.org/10.4304/tpls.4.7.1385-1394

Maleki, N. A. (2012). The effect of the involvement load hypothesis on improving Iranian EFL learners' incidental vocabulary acquisition in listening comprehension classes. *Australian Journal of Basic and Applied Sciences*, *6*(9), 119–128.

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

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). New York: Cambridge University Press.

Sarbazi, M.-R. (2014). Involvement load hypothesis: Recalling unfamiliar words meaning by adults across genders. *Procedia - Social and Behavioral Sciences*, *98*, 1686–1692. https://doi.org/10.1016/j.sbspro.2014.03.594

Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of involvement load, spacing, and intentionality. *TESL Canada Journal*, *34*(3), 140–164. https://doi.org/10.18806/tesl.v34i3.1277

Soleimani, H., & Rahmanian, M. (2014). The role of language glossing in a rooted theory: The involvement load hypothesis. *International Journal of Applied Linguistics & English Literature*, *3*(4), 6–13. https://doi.org/10.7575/aiac.ijalel.v.3n.4p.6

Soleimani, H., & Rahmanian, M. (2015). Visiting involvement load hypothesis and vocabulary acquisition in similar task types. *Theory and Practice in Language Studies*, *5*(9), 1883–1889. https://doi.org/10.17507/tpls.0509.16

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). Retrieved from http://link.springer.com/content/pdf/10.1057/9781137449788.pdf#page=140

Teng, F. (2015a). EFL vocabulary learning through reading BBC news: An analysis based on the involvement load hypothesis. *English as a Global Language*

Education (EaGLE) Journal, 1(2), 63–90. https://doi.org/10.6294/EaGLE.2015.0102.03

Teng, F. (2015b). Task effectiveness and vocabulary learning and retention in a foreign language. *Language Education and Acquisition Research Network (LEARN) Journal, 8*(1), 15–30.

Teng, F. (2017). The effects of task-induced involvement load on word learning and confidence judgments mediated by knowledge and regulation of cognition. *Educational Sciences: Theory & Practice, 17*(3), 791–808. https://doi.org/10.12738/estp.2017.3.0167

Tsubaki, M. (2012). *Vocabulary learning with graphic organizers in the EFL environment: Inquiry into the involvement load hypothesis* (Unpublished doctoral dissertation). Temple University, Philadelphia, PA.

Tu, H.-F. (2004). *Effects of task-induced involvement on incidental vocabulary learning in a second language* (Unpublished master's thesis, National Tsing Hua University). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.626.7187&rep=rep1&type=pdf

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*(3), 59–66. https://doi.org/10.5539/ijel.v4n3p59

Yang, S.-E. (2015). *The effects of the involvement load of tasks on English vocabulary learning of Korean high school students* (Unpublished master's thesis, Seoul National University). Retrieved from http://s-space.snu.ac.kr/bitstream/10371/127502/1/000000025064.pdf

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

Yaqubi, B., Rayati, R. A., & Gorgi, N. A. (2010). The involvement load hypothesis and vocabulary learning: The effect of task types and involvement index on L2 vocabulary acquisition. *Journal of Teaching Language Skills*, *29*(1), 145–163. https://doi.org/10.22099/jtls.2012.404

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*(1), 54–75. https://doi.org/10.1177/1362168816652418

Appendix H: Coding Scheme for Study 2

Coding column	Explanations of the column	Possible responses	Notes
study_no			
author			
year			
study			
exp			
participant_group			
publication_type		(1) journal - research journals (2) PhDthesis (3) MAthesis (4) bulletin - university journals (5) conference - conference	

	presentation; conference preceeding	
L1		
target_language		
institution	(1) elementary (2) secondary (3) university (4) language_school	Pre-university students in a certain language program were coded as language_school (Folse, 2006; Kim, 2008) When the research was carried out in a language institution even their institutional level is high school, it was coded as language institution. (i.e., Jahangiri & Abilipour2014)

activity	Type of activity		
activity2	Larger category of the type of activity	(1) fill_in (2) translation (3) writing (4) reading (5) graphic_organizers (6) matching (7) multiple_choice (8) speaking	(6) matching -any forms of matching activity (follow the authors' labeling: it does not matter what cognitive processes involved: coded just based on the format of an activity where participants were asked to match two items). (7) multiple_choice - any forms of multiple-choice activities (follow the authors' labeling: it does not matter what cognitive processes involved: coded just based on the format of an activity where participants were provided with multiple-choices and asked to select the most

			appropriate one). (8) speaking - oral-sentence-production (Hazrat, 2015) and retelling (Karalik & Merç, 2016)
need	a need component re-coded by the meta-analysts	0 - no need 1 - moderate need 2 - strong need	
search	a search component re-coded by the meta-analyst	0 - no search 1 - search was present	

evaluation	a evaluation component re-coded by the meta-analyst	0 - no evaluation 1 - moderate evaluation 2 - strong evaluation	
ILH	Total task-induced involvement load index re-coded by the meta-analyst		
evaluation_distinguishing_different_types_of_strong_evaluation	a evaluation component while distinguishing different types of strong evaluation	Evaluation component distinguishing different types of strone evaluation: 0 - no evaluation 1 - evaluation (ILH's moderate evaluation) 2 - sentence-level varied use 3 - composition-level varied use	Evaluation component including productive search: 0 - no evaluation 1 - evaluation: ILH's moderate evaluation was coded as evaluation 2 - sentence-level varied use: strong evaluation when each target word was used in a sentence (e.g., sentence writing, spoken sentence production)

			3 - composition-level varied use: strong evaluation when all target words were used in a composition (e.g., composition writing, retelling)
frequency	Number of times participants encountered or used the same set of target words		
mode	Which mode input was provided	(1) written, (2) spoken	Coded as spoken when participants were exposed to the target words in both spoken and written modes (e.g.,

			listening task with the provision of a glossary)
number_of_target_words	Number of target words participants were exposed in THAT treatment (i.e., learning condition) in question		
test_format		(1) meaning recall - e.g., translation (L2-> L1) (2) meaning recognition (3) form recall - e.g., translation (L1-> L2) (4) form recognition (5) form recognition - recognize whether target words were present in the treatment (6) VKS (7) gap-filling (8) productive use - when participants were asked to write	

		a sentence using a target word and the sentence was judged based on its semantic and grammatical accuracy; or just asked to use (Feng, 2015)
tests_max_score	Maximum score for the test	
pretest_mean		For VKS, 1 point x the number of target words was inserted when VKS's Category I scored 1 point.
pretest_test_SD		

how_many_days_until_the_immediate_test	Number of days between treatment and the immediate posttest
immediate_test_n	Number of participants who took the test in question
immediate_test_M	
immediate_test_SD	
how_many_days_until_the_delayed_test	Number of days between treatment and the delayed posttest
delayed_test_n	Number of participants who took the test in question
delayed_test_M	
delayed_test_SD	

immediate_test_ES	ES = $(posttest\ score - pretest\ score)/(test\ score\ maximum - pretest\ score)$
immediate_test_ES_SD	ES = (posttest score SD)/(test score maximum)
delayed_test_ES	ES = $(posttest\ score - pretest\ score)/(test\ score\ maximum - pretest\ score)$
delayed_test_ES_SD	ES = (posttest score SD)/(test score maximum)

Appendix I: Sensitivity Analysis for Study 2

In order to evaluate the robustness of the results, sensitivity analyses were carried out to investigate whether the results held when potential outliers were excluded from the analyses.

Following Viechtbauer and Cheung (2010) guidance and earlier meta-analyses (e.g., de Vos et al., 2018), we identified studies that were influencing the results significantly more than other studies by examining each study's Cook's distance and standardized difference of the beta (DFBETAS). Studies with Cook's distance higher than 0.85 and studies with a DEFBETAS value higher than 1 were identified as potential outliers.

When examining Cook's distance of the included studies on immediate posttests, Jing and Jianbin (2009), Martinez-Fernandez (2008), Teng (2015b), and Yang (2015) were identified as potential outliers. Similarly, on delayed posttests Jing and Jianbin (2009), Martinez-Fernandez (2008), Li (2014), Rott (2012), Soleimani and Rahmanian (2015), and Karalik and Merç (2016) were identified as potential outliers.

When examining DFBETAS, Jing and Jianbin (2009) and Martinez-Fernandez (2008) were identified as potential outliers. Similarly, Jing and Jianbin (2009), Li (2014), Rott (2012), and Martinez-Fernandez (2008) were identified as potential outliers on delayed posttests.

Because each study was independently conducted and included a different group of students and target words and varying learning conditions, studies identified as outliers do not necessarily mean the study is an outlier that does not reflect normal incidental vocabulary learning. Therefore, it is not appropriate to simply delete "outlier" studies from the analysis (see Hunter & Schmidt, 2004, for arguments about how to treat outliers). We followed Viechtbauer and Cheung's (2010) guidance and reran the whole analysis while excluding the studies identified as influential and compared the results to the results obtained when including all studies. The differences in the results revealed the

parts of the analysis that could be interpreted as robust and the parts of analysis that should be interpreted with caution.

The results that did not differ regardless of the inclusion of the outlier studies were (i) optimal ILH operationalization, (ii) Test format grouping, (iii) including Frequency only on immediate and Test day only on delayed posttests. The results that changed when excluding the outlier studies were that search and mode were included as meaningful predictors both on immediate and delayed posttests. The direction of the effect was the same when the outliers were included; that is, the inclusion of search negatively influenced learning, and spoken mode led to smaller learning gains than written mode.

Given that search was included on delayed posttests and mode was included on immediate posttests when analyzing all studies and their directions of influence were the same (i.e., the negative influence of search and disadvantage of spoken mode), the results from the sensitivity analysis point to the possibility that the negative influence of search and advantage of written mode will potentially be observed regardless of the timing of the test. However, given these results were only obtained when outlier studies were excluded, further research is warranted to draw a more definitive conclusion.

Furthermore, the explained variance was also examined when excluding the outliers. Table 1 showed the variance explained by ILH and the updated ILH (resulting model identified by analyzing the full dataset) at different levels (i.e., total variance, variance within-study levels) when outliers were excluded. Both on immediate and delayed posttests, the updated ILH led to greater explained variance than ILH.

Table 1. *The explained variance between the original ILH and the updated ILH*

	Immediate		Delayed	
	ILH	Updated ILH	ILH	Updated ILH

Total variance	17.9%	19.4%	3.8%	30.7%
Variance at within-study level	38.2%	76.8%	34.2%	69.7%

Notes. Immediate = Immediate posttests. Delayed = Delayed posttests.

References

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

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

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

Appendix J: Coding Scheme for Study 3

coding column	explanations of the column	possible responses	notes
study_no			
author			
year			
study			
exp			
participant_group			
publication_type		(1) journal - research journals (2) PhDthesis (3) MAthesis (4) bulletin - university journals (5) conference - conference	

		presentation; conference preceeding
L1		
target_language		
need (updated ILH components)	a need component re-coded by the meta-analysts	0 - no need 1 - moderate need 2 - strong need
search (updated ILH components)	a search component re-coded by the meta-analyst	0 - no search 1 - search was present

evaluation_distinguishing_different_types_of_strong_evaluation (updated ILH components)	a evaluation component while distinguishing different types of strong evaluation (updated ILH)	Evaluation component distinguishing different types of strone evaluation: 0 - no evaluation 1 - evaluation (ILH's moderate evaluation) 2 - sentence-level varied use 3 - composition-level varied use	Evaluation component including productive search: 0 - no evaluation 1 - evaluation: ILH's moderate evaluation was coded as evaluation 2 - sentence-level varied use: strong evaluation when each target word was used in a sentence (e.g., sentence writing, spoken sentence production) 3 - composition-level varied use: strong evaluation when all target words were used in a composition (e.g., composition writing, retelling)

activity - rough categorization	Larger category of the type of activity	(1) fill_in (2) translation (3) writing (4) reading (5) graphic_organizers_involving_ sentence_production (6) matching (7) multiple_choice (8) speaking	(5) graphic_organizers involving sentence production - Tsubaki's (2012) high involvement load condition was coded as graphic-organizers involving sentence production given the fact that the main activity was to write a sentence using each target word, while the low involvement load condition was coded as multiple-chioce becase its main activity was multiple choice. (6) matching -any forms of matching activity (follow the authors' labeling: it does not matter what cognitive processes involved: coded just based on the format of an activity where

	participants were asked to match two items). (7) multiple_choice - any forms of multiple-choice activities (follow the authors' labeling: it does not matter what cognitive processes involved: coded just based on the format of an activity where participants were provided with multiple-choices and asked to select the most appropriate one). (8) speaking - oral-sentence-production (Hazrat, 2015) and retelling (Karalik & Merç, 2016)

activity - fine-tuned categorization reflecting on the components of learning conditions	Type of activity	(1) glossed reading; (2) listening with a list of target words; (3) reading without reference to target words (4) glossed reading with comprehension questions requiring the understanding of target word; (5) glossed reading where target words were important for comprehension; (6) listening with a list of target words and comprehension questions requiring the understanding of target word; (7) reading in which target words were important for comprehension and dictionaries were provided; (8) reading with the support of dictionaries plus	(1) - (3) : no components (4)-(8) : moderate need (9) - (16) : moderate need, evaluation (17) - (19) : moderate need, sentence-level varied use (20) - (22) : moderate need, composition-level varied use Reading conditions that included need but not clearly explained how need was ecilitated were coded as either (5) glossed reading where target words were important for comprehension or (7) reading in which target words were important for comprehension and dictionaries were provided,

	comprehension questions requiring the understanding of target word	based on their reference type (glosses or dictionary).
	(9) fill-in-the-blanks in passages; (10) reading with multiple-choice glosses; (11) fill-in-the-blanks in sentences; (12) multiple-choice questions; (13) translation; (14) matching; (15) reading with dictionaries (multiple meanings were presented for each target word and participants needed to determine the meaning that fit the context); (16) sentence-combinations, where participants combine segments of a sentence to regenerate the sentence	Reading conditions that included evaluation where dictionaries were provided but not clearly explained how evaluation was elicited were also coded as (15) reading with dictionaries (multiple meanings were presented for each target word and participants needed to determine the meaning that fit the context).

	(17) sentence writing; (18) graphic organizers involving sentence-production; (19) oral sentence-production
	(20) composition writing; (21) retelling; (22) summary writing

frequency	Number of times participants encountered or used the same set of target words		
mode	Which mode input was provided	(1) written, (2) spoken	Coded as spoken when participants were exposed to the target words in both spoken and written modes (e.g., listening task with the provision of a glossary)
number_of_target_words	Number of target words participants were exposed in THAT treatment (i.e., learning condition) in question		

test_format		(1) meaning recall - e.g., translation (L2-> L1) (2) meaning recognition (3) form recall - e.g., translation (L1-> L2) (4) form recognition (5) form recognition - recognize whether target words were present in the treatment (6) VKS (7) gap-filling (8) productive use - when participants were asked to write a sentence using a target word and the sentence was judged based on its semantic and grammatical accuracy; or just asked to use (Feng, 2015)
tests_max_score	Maximum score for the test	

pretest_mean		For VKS, 1 point x the number of target words was inserted when VKS's Category I scored 1 point.
pretest_test_SD		
how_many_days_until_the_immediate_test	Number of days between treatment and the immediate posttest	
immediate_test_n	Number of participants who took the test in question	
immediate_test_M		
immediate_test_SD		
how_many_days_until_the_delayed_test	Number of days between treatment and the delayed posttest	

delayed_test_n	Number of participants who took the test in question
delayed_test_M	
delayed_test_SD	
immediate_test_ES	ES =$(posttest\ score - pretest\ score)/(test\ score\ maximum - pretest\ score)$
immediate_test_ES_SD	ES = (posttest score SD)/(test score maximum)
delayed_test_ES	ES = $(posttest\ score - pretest\ score)/(test\ score\ maximum - pretest\ score)$
delayed_test_ES_SD	ES = (posttest score SD)/(test score maximum)

Appendix K: Details of the Results Including all Predictor Variables

The complete results of the meta-regression models on immediate and delayed posttests are presented in Tables 1 and 2, respectively. While the results presented in the main text (Table 2) are the results of the meta-regression models without intercept (a.k.a. intercept only models), the results provided here are those with intercept. The predictor variables that are not directly related to activity types (i.e., search, frequency, mode, test format, and test day) were not presented in the results in the main texts but included in the meta-regression models as covariate, and their results correspond to those reported here because the results of covariates do not change regardless of whether a model includes intercept or not.

Table 1

The Results Including all Predictor Variables on the Immediate Posttest

		95% CI		
Predictor variables	Estimate	Lower	Upper	*p*
Intercept	0.070	-0.093	0.232	.377
Need for comprehension	0.213	0.032	0.393	.028
Evaluation	0.084	0.041	0.128	.001
Sentence-level varied use	0.153	0.080	0.225	< .001
Composition-level varied use	0.233	0.129	0.337	< .001
Frequency	0.095	0.016	0.175	.028
Mode: Spoken	-0.098	-0.226	0.030	.093

Test: Productive recall	-0.125	-0.221	-0.030	.022
Test: Recognition	0.227	0.028	0.426	.040
Test: Other	-0.099	-0.158	-0.040	.009
Total explained variance	.171			
Between-study variance explained	.000			
Within-study variance explained	.594			

Note. Need for comprehension refers to the need to understand target words while reading. 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 2

The Results Including all Predictor Variables on the Delayed Posttest

		95% CI		
Predictor variables	Estimate	Lower	Upper	*p*
Intercept	0.187	0.063	0.310	.007
Need for comprehension	0.140	0.028	0.252	.022
Search	-0.049	-0.120	0.021	.149

Evaluation	0.090	0.043	0.138	.001
Sentence-level varied use	0.113	0.060	0.166	< .001
Composition-level varied use	0.208	0.155	0.261	< .001
Test day	-0.004	-0.007	-0.001	.015
Test: Productive recall	-0.120	-0.272	0.031	.092
Test: Recognition	0.216	0.049	0.383	.032
Test: Other	-0.087	-0.128	-0.046	.004
Total explained variance	.344			
Between-study variance explained	.186			
Within-study variance explained	.604			

Note. Need for comprehension refers to the need to understand target words while reading. 95% CIs and *p*-values were calculated based on the robust Variance Estimation. For reference level, Test format was set as receptive recall.

Curriculum Vitae

Name:	Akifumi Yanagisawa
Post-secondary Education and Degrees:	Shinshu University Nagano, Nagano, Japan 20009-2013 B.A. Shinshu University Nagano, Nagano, Japan 2013-2016 M.A. The University of Western Ontario London, Ontario, Canada 2016-2020 Ph.D.
Honours and Awards:	SLRF 2019 Student Travel Award SLRF Organizing Committee 2019 Graduate Excellence in Teaching Award The University of Western Ontario 2020
Related Work Experience	Part-time Lecturer The University of Western Ontario 2017, 2019

Research Assistant

The University of Western Ontario

2016-2020

Publications:

Webb, S., **Yanagisawa, A**., & Uchihara, T. (in press). How effective are intentional vocabulary learning activities? A meta-analysis. *The Modern Language Journal.*

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:10.1017/S0272263119000688

Yanagisawa, A., & Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.) *The Routledge Handbook of Vocabulary Studies* (pp. 371-386). New York: Routledge. https://doi.org/10.4324/9780429291586-24

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

Uchihara, T. & **Yanagisawa, A.** (2017). Lessons from Western's symposium on teaching and learning vocabulary in another language. *Contact*, 41(1), 15-22.

Yanagisawa, A. (2016). *Effects of Receptive and Productive Retrieval Practice on L2 Vocabulary Knowledge: Learning Pseudo-English Words Paired with Hand Movements* (Unpublished MA Thesis). Shinshu University, Nagano, Japan.

Yanagisawa, A. (2016). The effects of receptive and productive word retrieval practice on second language vocabulary learning. *KATE Journal*, 30, 139-152. https://doi.org/10.20806/katejournal.30.0_139