EFFECTS OF TASKS AND GLOSSES ON L2 INCIDENTAL VOCABULARY LEARNING: META-ANALYSES

A Dissertation

by

SHU-FEN HUANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Curriculum and Instruction

EFFECTS OF TASKS AND GLOSSES ON L2 INCIDENTAL VOCABULARY LEARNING: META-ANALYSES

A Dissertation

by

SHU-FEN HUANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Zohreh R. Eslami
Committee Members,	Victor Willson
	Lynn M. Burlbaw
	Janet Hammer
Head of Department,	Dennis Smith

August 2010

Major Subject: Curriculum and Instruction

ABSTRACT

Effects of Tasks and Glosses on L2 Incidental Vocabulary Learning:

Meta-analyses. (August 2010)

Shu-fen Huang, B.A., Chaoyang University of Technology;

M.Ed., Texas A&M University

Chair of Advisory Committee: Dr. Zohreh R. Eslami

This study investigated the effects of output stimulus tasks and glosses on L2 incidental vocabulary learning. Two meta-analytic studies were conducted. The first was intended to provide a systematic statistical synthesis of the effects of output stimulus tasks on L2 incidental vocabulary learning. A total of 12 studies were included in this meta-analysis. Results showed that language learners gained more benefit from using output stimulus tasks to learn vocabulary than those who only read a text. Results also supported the involvement load hypothesis that language learners who perform a task with a higher extent of involvement load gain more L2 vocabulary. As opposed to studies with the low level of design quality, studies with high and medium levels of design quality were more likely to detect statistically significant differences among groups with different output stimulus tasks. Moreover, results suggested that time on task had a positive impact on L2 vocabulary learning. Learners who read a combination of expository and narrative texts outperformed those who only read either an expository or a narrative text in the vocabulary posttest. Learners who read a text with text-target

word ratios of less than or equal to 2% did not learn significantly more vocabulary than those who read a text with a ratio of 2% to 5%.

The second meta-analysis study used meta-analytic techniques to explore the effects of L1 textual and image-based glosses on second language (L2) incidental vocabulary learning while reading. Results revealed that language learners who were provided with textual glosses gained more vocabulary than those who had no access to glosses. Results suggested that text-target word ratios played an important role in second language vocabulary learning. Language learners who read a passage with a text-target word ratio of ≤2% outperformed those who read a passage with a text-target word ratio between 2% and 5%. No statistically significant difference was found between the groups that were provided with multiple-choice and single glosses. Compared to paper-and-pencil environments, computer-assisted settings did not significantly enhance L2 vocabulary learning. Language learners who read narrative reading materials did not significantly outperform those who were exposed to expository texts with regard to incidental vocabulary learning. No significant difference in L2 vocabulary learning was observed between groups who were given L1 textual glosses and those who had access to L1 textual+image-based glosses.

DEDICATION

To my family

ACKNOWLEDGEMENTS

Writing a dissertation is a tough undertaking. I would like to thank all of my dissertation committee members for their time, insight, and contribution on my intellectual growth. In particular, I would like to thank Dr. Zohreh R. Eslami, my mentor and committee chair. This dissertation could not have been completed without her expertise and constant guidance. Dr. Eslami encouraged me to work to her standards and gave me her extensive support. She helped me work through the difficulty of my doctoral program, grow intellectually, and develop professionally as a scholar. In addition, she taught me how to read papers critically and develop the ability to write clearly. Her tireless commitment and attention to detail influenced me to create a work that I am proud of.

Moreover, I would like to express my deep appreciation to the other three members of my committee: Dr. Victor Willson, Dr. Lynn M. Burlbaw, and Dr. Janet Hammer. Special thanks to Dr. Willson for helping me develop statistical knowledge, and appreciate insightful advice on data analysis. I also thank Dr. Burlbaw and Dr. Hammer, whose probing questions made me think deeply and critically. I gratefully thank them for graciously contributing their time and helpful input that shaped this dissertation into a better work.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Vocabulary is pivotal in second language learning. Nation (2006) suggests that a language learner needs a vocabulary of 6,000 to 7,000 word families to understand a typical movie, and 8,000 to 9,000 word families to comprehend written text. This requirement makes language acquisition impossible solely through explicit language learning. Thus, incidental vocabulary learning serves as an important method to help language learners increase their vocabularies.

Studies (Day, 1991; Day & Swan, 1998; Pitts et al., 1989) have shown that learning vocabulary incidentally through reading is effective. However, in a synthesis review of L2 vocabulary learning, Schmitt (2008) notes that incidental vocabulary learning during reading has a low pick-up rate. Horst et al. (1998), for example, found that in incidental learning, only about one word out of every 12 was accurately identified. Therefore, incidental vocabulary learning via reading alone may not be sufficient.

Much research has focused on how to enhance the effectiveness of incidental vocabulary learning in reading by using stimulus techniques such as output tasks, textual glosses, and think-aloud activities (Hill & Laufer, 2003; Kim, 2008; Ko, 1995; Lee, 1995; Min, 2008; Rott, 2004; Watanabe, 1997). Among these studies, two types of stimulus techniques—output tasks and textual glosses—have been widely used to

---

This dissertation follows the style of *The Modern Language Journal.*

enhance L2 incidental vocabulary acquisition. Some studies (Hill & Laufer, 2003; Kim, 2008; Min, 2008; Rott, 2004) employed tasks that required learners to produce output alongside their reading. Rott, Williams, and Cameron (2002) contend that greater emphasis on new vocabulary can lead to a greater likelihood that vocabulary will be acquired and retained. Learning and retention are improved when learners use, reformulate, or elaborate on this new information, because these processes induce connections between existing and new knowledge (Crailk & Tulving, 1975). Compared to input, output requires more mental effort on the part of learners. Learners can pretend to comprehend while reading, but they cannot do so while speaking or writing. The process of vocabulary growth may reach a plateau unless learners are given the opportunity to develop skills in its use (Nation, 2001).

Other studies (Ko, 1995; Lee, 1995; Watanabe, 1997) used textual glosses to reduce inaccurate guessing and facilitate vocabulary learning by enhancing learners' noticing of unknown words in text. Researchers (Hulstijn, 1992; Hulstijn et al., 1996; Watanabe, 1997) have also found that glosses facilitate second-language learners' vocabulary growth. Research has further indicated that language learners with access to marginal vocabulary glosses demonstrate greater vocabulary growth than those without glosses (Hulstijn, 1992; Hulstijn, et al., 1996; Jacob, Dufon, & Fong, 1994; Watanabe, 1997).

In addition to textual glosses, image-based glosses are often used. Researchers (Al-Seghayer, 2001; Chun & Plass, 1996; Kost, Foss, & Lenzini, 1999; Yoshii & Flaitz, 2002) investigated the combined effect of L1 and image-based glosses on L2 vocabulary

learning. Chun and Plass (1996) found that students with access to textual and image-based glosses gained greater L2 vocabulary than those who were provided textual glosses only. However, Al-Seghayer (2001) found that students showed greater word gain if the words were provided with textual glosses and video clips, compared to students with access only to pictorial glosses.

Purpose of the Study

Research results on the effect of different output stimulus tasks and glosses on L2 vocabulary learning have not been conclusive. A comprehensive account of the effectiveness of these two techniques on language learners' vocabulary acquisition is warranted to improve the knowledge of L2 incidental vocabulary learning. A meta-analysis provides a systematic approach to characterizing patterns in quantitative studies, which could be difficult to detect using a narrative approach. A meta-analysis offers an overall picture of a related topic across research studies and helps researchers identify effective treatments within a domain.

Only one meta-analysis (Wa-Mbaleka, 2006) investigated the effectiveness of depth of processing in reading activities on L2 vocabulary learning during reading. In Wa-Mbaleka's (2006) study, the effects of Depth of Processing on L2 vocabulary learning were analyzed within eight levels: "Level 0:control group, no treatment; Level 1: reading authentic text without dictionaries or glosses; Level 2: reading modified texts from class textbooks or reading plus other L2 activities; Level 3: reading with target words highlighted; Level 4: reading with glosses or dictionaries; Level 5: reading + some productive ways of using new words; Level 6: writing; Level 7: reading + direct

learning of L2 words; and Level 8: direct instruction/learning of L2 words" (Wa-Mbaleka, 2006, p.207).

No differences were detected across these eight levels. Wa-Mbaleka (2006) claimed that the results were "too heterogeneous to allow meaningful discussion" (p.208), noting that this variability may be due to too many levels of processing (eight). Wa-Mbaleka (2006) suggested that a simpler classification was warranted to show meaningful results. Moreover, moderators, such as genres of reading materials (e.g., narrative or expository) and text-target word ratios, were not investigated in Wa-Mbaleka's (2006) study. The current meta-analysis attempts to shed new light on this issue by separating the effect of output tasks and glosses on L2 incidental vocabulary learning during reading, and incorporating the relevant moderators in the design.

The following research questions will be addressed in Chapter II:

1. What are the overall effects of different output stimulus tasks on L2 incidental vocabulary learning?
2. Does the design quality of the study (i.e. low, medium, or high design quality) have different effects on detecting statistically significant differences among the groups with different output stimulus tasks (fill-in-the-blank, sentence writing, composition, or a combination of these tasks)?
3. Which type of tasks (i.e. fill-in-the-blank, sentence writing, composition, or a combination) has the largest effect on L2 incidental vocabulary learning?
4. Does the type of text (genre) (narrative and expository texts) have different effects on L2 incidental vocabulary learning?

5. Do text-target word ratios (i.e. ≤2%, between 2% and 5%, and ≥5 %) have varying effects on L2 incidental vocabulary learning?
6. Does the treatment length have different effects on L2 incidental vocabulary learning?

Glosses are another common technique to foster vocabulary learning. Different types of glosses (single textual, multiple-choice, and image-based) are used to prevent learners from inferring incorrect meaning of unknown words in text. Glosses also direct learners' attention to target words, leading to greater vocabulary acquisition (Nation, 2001). Studies have shown that language learners who were provided with textual glosses gained more vocabulary than those who had no access to any types of glosses (Bowles, 2004; Cheng & Good, 2009; Jacobs, et al., 1994; Ko, 1995; Hulstijn et al., 1996). However, the inferred meaning of an unknown word is often better retained than the given meaning (Nation, 2001); providing the meaning of a word does not allow learners opportunities to derive the meaning of unknown words in text. On the other hand, context richness varies with each text. Laufer (2005) mentioned that linguistics cues for the meaning of unknown words are often unavailable or misleading. In addition, the words that are included in the cues are sometimes unknown to the learners. As a result, learners cannot make effective use of all linguistics cues; insufficient information in text presents a challenge to learners trying to infer the meaning of unfamiliar words. Hulstijn (1992) proposed the use of multiple-choice glosses as a remedy. Multiple-choice glosses combine the advantages of both single textual glosses and inference. Multiple-choice glosses decrease the possibility of guessing incorrect word meanings and reduce the challenges caused by insufficient information in text. However, studies

showed mixed results pertaining to the effects of multiple-choice glosses in L2 incidental vocabulary learning. Some studies (Nagata, 1999; Rott, 2005) showed that language learners who were provided with multiple-choice glosses performed better than those who had access to single textual glosses, while other research (Watanabe, 1997) found no significant difference between the effectiveness of these two types of glosses.The use of image-based glosses is also widespread in L2 vocabulary learning. Images or visual presentations of the target words strengthen the link between the mental image and semantic meaning of a word. Research (Plass, Chun, Mayer, & Leutner, 2003; Yanguas, 2009) has shown that learners who were provided with imaged-based glosses outperform those who were not given any glosses. Studies have also found (Akbulut, 2007; Chun & Plass, 1996; Kost, Foss, & Lenzini, 1999;Yanguas, 2009;Yoshii & Flaitz, 2002) that the combination of textual and pictorial glosses led to greater vocabulary gains than textual glosses only. To summarize, glosses can facilitate second language vocabulary development; different types of glosses have various effects on L2 vocabulary learning.

A recent meta-analysis (Abraham, 2008) examined the effect of computer-mediated glosses on L2 incidental vocabulary learning by comparing six studies. Abraham (2008) reported that computer-mediated glosses have an overall effect size of 1.40 for the immediate vocabulary posttest, and 1.25 overall effect size for the delayed vocabulary posttest. Regarding the effect of glosses on learners with different proficiency levels, the results showed that learners with an intermediate proficiency level benefited the most, and beginning learners, the least. However, the difference was not

statistically different. Although Abraham's (2008) meta-analysis sheds light on the effects of textual glosses on L2 incidental vocabulary learning, some questions remain unanswered. Abraham (2008) focused only on studies conducted in computer-assisted environments. The magnitude of the effects of different textual glosses on L2 vocabulary learning conducted in paper-and-pencil settings might be different. In addition, many moderator variables related to vocabulary learning, such as the ratio of text-target word and multiple-choice glosses, were not investigated in this study. A meta-analytic study is warranted to address the effects of these two types of glosses on L2 incidental vocabulary learning. Therefore, the current research aims to address the following research questions in Chapter III:

1. What are the overall effects of L1 textual glosses and image-based glosses on L2 incidental vocabulary learning?
2. Does the design quality of the study (low, medium, or high design quality) have different effects on detecting statistically significant differences among the groups with different glosses (L1 textual glosses, L1 textual + imaged-based glosses, or no glosses)?
3. Do multiple-choice glosses and single glosses cause different effects on L2 incidental vocabulary learning with the different types of glosses?
4. Does the type of setting (computer-assisted or paper-and-pencil based) have different effects on L2 incidental vocabulary learning with the different types of glosses?
5. Do the text-target word ratios (≤2%, between 2% and 5%, and ≥5 %) have different effects on L2 incidental vocabulary learning with different types of glosses?

6. Does the genre of reading text (narrative and expository texts) have different effects on L2 incidental vocabulary learning with different types of glosses?

Delimitations

The difficulty of article identification and lack of comprehensiveness is a delimitation in this study. Most studies included in the current investigation were found using electronic searchable databases of scholarly literature. Therefore, some journals or book chapters are not included in these databases. In the same vein, dissertations and theses may not be available within the year that their authors graduated.

Definitions of Terms

Meta-analysis: A meta-analysis uses a systematic synthesis of research studies that yield quantitative findings. A meta-analysis has three advantages. First, it provides research findings in a sophisticated fashion, which differs from findings represented in statistical significance. Second, it is able to detect effects that are obscure in narrative summaries of findings. Third, it provides a systematic approach to analyzing information from a large number of research findings (Lipsey & Wilson, 2001).

Incidental vocabulary learning: Incidental vocabulary learning is defined as learning vocabulary items as a by-product of language use while language-learners' attention is focused on the meaning to be conveyed. Incidental vocabulary learning occurs when language learners read a text with the intention of comprehension.

L1 glosses: Vocabulary is annotated with one or more synonyms or is provided with definitions or explanations in the participants' first language. In the present study, L1 glosses are defined as an instant look-up capability: the provision of a dictionary,

synonyms, or definitions. In other words, language learners have access to the meaning of an unknown word in their native language.

Word family: A word family consists of a headword, its affixes (such as -ly, -ness, and un-) and derivations. For example, *happy*, *happiness*, *happily*, and *unhappy* are words from one word family.

Image-based glosses: These are vocabularies that are glossed with visual representations, such as a picture, animation, or video clip.

Outline of the Dissertation

This dissertation contains four chapters. Chapter I provides an overview of the purpose of the current study, the statement of the problem, the definitions of important concepts, and delimitations. Chapter II is a meta-analysis study on the effect of different output stimulus tasks on L2 incidental vocabulary learning. Chapter III is another meta-analysis, this one focusing on the effect of different types of glosses on L2 incidental vocabulary learning. Chapter IV contains the conclusion of findings, pedagogical implications, limitations of the study, and recommendations for further study.

CHAPTER II

THE EFFECTS OF TASK-INVOLVEMENT LOAD ON L2 INCIDENTAL VOCABULARY LEARNING: A META-ANALYTIC STUDY

Overview

This meta-analytic study was intended to provide a systematic statistical synthesis of the effects of output stimulus tasks on L2 incidental vocabulary learning. A total of 12 studies were included in this meta-analysis. Five mediator variables were examined: design quality, types of output stimulus task, time on task, genres of text, and text-target word ratios. Results showed that language learners who completed an output stimulus task outperformed those who only read a text. Results also supported the involvement load hypothesis; language learners who performed a task with a higher degree of involvement load gained more L2 vocabulary. Studies with high and medium levels of design quality were more likely to detect statistically significant differences among groups with different output stimulus tasks, compared to studies with the low level of design quality. Furthermore, results indicated that time on task had positive effects on L2 vocabulary learning. Learners who read a combination of expository and narrative text gained more vocabulary than those who only read either an expository or narrative text. Learners who read a text with text-target word ratios of less than or equal to 2% did not learn significantly more vocabulary than those who read a text with a ratio of 2% to 5%. Suggestions for further studies are provided.

Introduction

Incidental vocabulary learning has long been a focus in the field of learning English as a second or foreign language. Research (Day, 1991; Day & Swan, 1998; Pitts et al., 1989) has shown that L2 incidental vocabulary learning through reading is a viable approach. However, incidental vocabulary learning alone does not have a high pick-up rate (Schmitt, 2008). Vocabulary retention is partly determined by how information is processed (Crailk & Lockhart, 1972). Compared to less mentally elaborated lexical information, more mentally elaborated information leads to greater retention (Hulstijn & Trompetter, 1998). Many stimulus techniques—glossing, bolding, italicizing, color-coding, or word-focused exercises—have been used to increase the effectiveness of incidental vocabulary learning while reading (Hill & Laufer, 2003; Kim, 2008; Ko, 1995; Lee, 1995; Min, 2008; Rott, 2004; Watanabe, 1997). Among these techniques, output stimulus tasks—such as fill-in-the-blank, sentence writing, and composition writing—are commonly used to make language learners process unknown words on a deeper level and facilitate vocabulary learning.

Although output stimulus tasks have been widely used in the field of language learning, only one meta-analytic study (Wa-Mbaleka, 2006) has investigated their effects on L2 incidental vocabulary learning. Although Wa-Mbaleka (2006) investigated various levels of depth of processing, such as the use of a glossary, highlighted target words, writing, and explicit teaching, the categorization of levels did not produce a significant difference. Wa-Mbaleka (2006) claimed that the results were “too heterogeneous to allow meaningful discussion” (p.208), noting that this variability may

be due to too many levels of processing (eight). Wa-Mbaleka (2006) suggested that a simpler classification was warranted to show meaningful results. Moreover, moderators, such as genres of reading materials (e.g., narrative or expository) and text-target word ratios, were not investigated in Wa-Mbaleka's (2006) study. The current meta-analysis attempts to shed new light on this issue by separating the effect of output tasks and glosses on L2 incidental vocabulary learning during reading, and incorporating relevant moderators in the design.

Literature Review

This section provides an overview of previous studies of L2 incidental vocabulary learning that have used reading and output stimulus task. The overview focuses on three aspects: the theoretical framework behind output stimulus task, the effects of the output stimulus tasks on incidental vocabulary learning, and previous meta-analysis related to output tasks and vocabulary learning.

Theoretical Framework

Research on output stimulus tasks is guided by the theoretical perspectives of depth of processing, output and noticing hypothesis, and the involvement load hypothesis.

Depth of Processing

How much attention a learner pays to a given word is highly related to the extent to which he or she will remember that word. Researchers widely agree with Crailk and Lockhart's (1972) idea of depth of processing, "where greater depth implies a greater degree of semantic or cognitive analysis" (p.675). This theory contends that deeper

retrieval processes have two distinct long-term advantages over less-complex processes. First, more-complex retrieval processes reactivate or strengthen the encodings of an item, which are less susceptible to interference, thus making retrieval more conducive to long-term retention. Second, more-complex retrieval processes take a slower and more complicated route to stored memory, making subsequent retrieval easier. Learning and retention are improved when learners use, reformulate, or elaborate on this new information, because these processes induce connections between extant and new knowledge (Crailk & Tulving, 1975). The likelihood that new information will be stored in long-term memory is largely determined by the extent to which the information is originally processed. Similarly, Rott, Williams, and Cameron (2002) contend that greater attention to new vocabulary can lead to a greater likelihood that vocabulary will be acquired and retained. Thus, tasks that involve a deep level of processing of new words better facilitate learning.

Output and Noticing Hypothesis

Comprehensible input alone is not sufficient for language learning. Many researchers now believe that output also is an important component of language learning. In addition, noticing (or an awareness of linguistics features), is a prerequisite for the process of language learning (Schmidt, 1994). Learners do not acquire meaning, syntax, or any other element of language use unless they are aware of them. Noticing and output hypothesis have emerged as an answer to the insufficiency of comprehensible input. Gass (1988) proposed that integration of lexical elements into a language learner's mental lexicon goes through five stages: apperceived input, comprehended input, intake,

integration, and output. In the apperceived input stage, data are comprehended by language learners. However, not all data are noticed by language learners; lexical items with highly frequent occurrences or those that have a connection to prior knowledge are more likely to be noticed from the available sensory information. The second stage is comprehended input. Unlike Krashen(1982), Gass (1988) claims that comprehended input is a multi-staged concept. According to Gass (1988), comprehended input has multiple levels of comprehension, such as general meaning, semantics, and detailed structural analyses. Language learners might understand parts of an utterance and grasp its syntactic or phonological pattern. Not all comprehended input will go through to the next stage: intake. Intake is a process of attempted integration of linguistic information, mediating between target language input and the learners' existing internalized rules. The following stage is integration, which is an outcome of the intake process. Integration involves changes to the internalized second language rule system based on the new information. Output, the final stage, plays a large role in the conversion of comprehended input into intake, and learners' output, which can represent the integrated knowledge. In addition, output forces learners to reflect upon the adequacy of the specific target language knowledge for their intended messages.

Similarly, Swain (1995) contends that output serves a noticing function in language learning. Output production helps language learners who attempt to produce the target language to notice gaps in their language proficiency. It also helps learners notice the discrepancy between what they intended to say but could not say precisely or completely. When learners notice the gap and realize they do not know the word in the

L2, they return to the text with more focused attention and fill the newly discovered gap. Compared to input, output requires more mental effort on the part of learners because they alone control their vocabulary acquisition. Learners can pretend to comprehend while reading, but they cannot do so while speaking or writing. The process of vocabulary growth may reach a plateau unless learners are given the opportunity to put this vocabulary to use and develop skills in its use (Nation, 2001). Even if learners have a sufficiently large vocabulary, further vocabulary growth will not be achieved.

Involvement Load Hypothesis

The Involvement Load Hypothesis proposed by Hulstijn and Laufer (2001) highlights the importance of the motivational-cognitive construct of involvement, which has three elements: need, search, and evaluation. Need is the motivational, noncognitive dimension of involvement that exists for language learners when an unknown word is required to finish a given task. Need is considered moderate when it is imposed by an external agent and strong when learners are intrinsically motivated. For instance, need is moderate in the case of a teacher asking for the use of a word in a sentence. On the other hand, need is strong when self-imposed by learners who look up the meaning of an unknown word in a dictionary while writing a composition.

Search and evaluation are referred to as the cognitive aspects of involvement; both require focusing on word forms and meaning. Search is present when learners are required to identify the meaning of an unknown word in a dictionary or by consulting teachers or students, whereas search is absent when this effort is not required, such as when an unknown word's definition is glossed.

Evaluation is a decision-making process during tasks, such as "a comparison of a given word with other words, a specific meaning of a word with its other meaning, or comparing the word with other words in order to assess if the word does or does not fit its context" (Hulstijn & Laufer, 2001, p.544). Evaluation is moderate when a task requires learners to acknowledge differences between words provided in a given context (e.g., fill-in-the-blank tasks) and strong when the task requires learners to decide the meaning of unknown words and then combine these words with known words in the original context (e.g., sentence writing and composition).

Not all tasks require learners to need, search for, and evaluate the meaning of unknown words. A task in which these elements are stronger has a higher level of involvement load. Language learners who are engaged in a task with a higher level of involvement load are more likely to retain vocabulary.

Previous Research on Output Stimulus Tasks

Many studies (Beal, 2007; Hulstijn & Trompetter, 1998; Hulstijn & Laufer, 2001) have explored the effects of output stimulus tasks on incidental L2 vocabulary acquisition while reading. Hulstijn and Trompetter (1998) examined the effects of a writing task on incidental vocabulary learning. Participants were high school students in the Netherlands who were learning French as a second language. They were assigned to either the reading only or writing groups. In the reading group, participants were required to read a text of a French weather report. After reading the same text as participants in the reading-only group, participants in the writing group were asked to write a weather report in French using ten target words. An unexpected vocabulary test

was given to both groups after the experiment. Results revealed that participants in the writing group outperformed those in the reading-only group.

Hulstijn and Laufer's (2001) study further explored the effects of output stimulus tasks on foreign language vocabulary by using various tasks with different degrees of involvement load. Participants who were advanced EFL university students were randomly assigned to one of three groups: fill-in-the-blank, composition, and reading comprehension with marginal glossing. All groups were given the same text to read for comprehension, with the target words included. The reading comprehension with marginal glossing group received the text that included the definitions of the target words in the margin. Participants in the fill-in-the-blank group read the text and also were asked to complete gapped sentences with the target words. The composition group read the text and also wrote a composition with the target words. Time on task varied due to the nature of the tasks. The composition group received the most time to complete the task, followed by the fill-in-the blank group and the reading comprehension with marginal glossing. An immediate posttest was given without advance notice to all participants after the experiment. Results showed that the composition group earned the highest score. However, the fill-in-the-blank group did not produce significantly higher retention than the reading comprehension with marginal glossing group. Results partially supported the hypothesis that words processed with a higher involvement load are retained better.

Tu (2004) explored the effect of task involvement load on L2 vocabulary learning and also investigated whether time on task had an impact on vocabulary

learning. Two experiments were conducted. The research designs of these two experiments were identical, except for the time-on-task factor. In the first experiment, time on task was not controlled; students were allowed to take as much time as they needed to complete their task. In the second experiment, time on task was held constant across groups. Participants were 267 Taiwanese high school students from six classes. Each class was randomly assigned to one of three conditions: reading a text, reading comprehension plus fill-in-the-blank activity, and writing a composition using target words. Results of the first experiment showed that the composition writing group performed best, followed by the fill-in-the-blank group, then the reading group. Results of the second experiment partially supported the involvement load hypothesis. Although the composition writing group gained the most vocabulary among the groups and the fill-in-the-blank group outperformed the reading group, the results were not all different in terms of statistical significance. These findings suggested that the effect of task type might be influenced by other factors when the amount of time on task was held constant. In other words, when learners were exposed to new words for an equal amount of time while performing various tasks, the construct of task involvement load might not be the only factor determining the effect of vocabulary learning.

Similarly, Beal (2007) investigated the effect of involvement load on foreign language reading and vocabulary acquisitions. A total of 118 high-intermediate ESL students at a Canadian college participated in the study. They were assigned to one of the following groups: textual glosses, multiple-choice glosses, sentence production, and a control group, which was asked only to read a text. In addition to reading a text, each

group except for the control group was instructed to perform a task. Those in the textual glosses group engaged in highlighting activities and were given a list of target words with their textual glosses. The multiple-choice glosses group had to select the correct meaning for target words from multiple-choice glosses of synonyms or definitions. Members of the sentence production group were asked to define and write sentences using target words. An unexpected, immediate posttest was administrated to all participants. Results showed that participants in the sentence production group performed the best, followed by the multiple-choice glosses, textual glosses, and control groups. These results support the effects of output task stimulus. The groups that were required to complete an additional task performed better on the posttest than those that did not. Furthermore, results were in accordance with the Involvement Load hypothesis; participants in a higher-involvement load group scored better than groups with a lower-involvement load.

In summary, we can state that output stimulus tasks facilitate incidental vocabulary learning. Involvement load plays a role in determining the extent to which language learners acquire second-language vocabulary. The above-mentioned studies suggest that the presence of output stimulus tasks and greater extent of the involvement load were beneficial in second-language vocabulary learning.

Previous Meta-analysis on Incidental Vocabulary Learning

Using a meta-analysis, Wa-Mbaleka(2006) investigated the effectiveness of depth of processing in reading activities on L2 vocabulary learning during reading. Wa-Mbaleka (2006) studied the effects of depth of processing on L2 vocabulary learning

within eight levels: “Level 0:control group, [reading treatment only]; Level 1: reading authentic text without dictionaries or glosses; Level 2: reading modified texts from class textbooks or reading plus other L2 activities, [such as discussion, think-aloud, or recall activities for target words]; Level 3: reading with target words highlighted; Level 4: reading with glosses or dictionaries; Level 5: reading+some productive ways of using new words; Level 6: writing; Level 7: reading+direct learning of L2 words; and Level 8: direct instruction/ learning of L2 words” (Wa-Mbaleka, 2006, p.207). No differences were detected across these eight levels. Wa-Mbaleka (2006) claimed that the results were “too heterogeneous to allow meaningful discussion” (p.208), noting that this variability may be due to too many levels of processing (eight levels). Wa-Mbaleka (2006) suggested that a simpler classification was warranted to show meaningful results. Moreover, this study did not investigate moderators, such as the genres of reading materials (e.g., narrative or expository) and the text-target word ratios. The current meta-analysis attempts to shed new light on this issue by investigating the effects of output tasks L2 incidental vocabulary learning during reading and incorporating the relevant moderators in the design.

Research Questions

The present meta-analysis was designed to answer the following research questions (RQs):

1. What are the overall effects of different output stimulus tasks (e.g., fill-in-the-blank, sentence writing, composition, or a combination of these tasks) on L2 incidental vocabulary learning?

2. Does the design quality of the study (i.e., low, medium, or high levels of design quality) have varying effects of detecting statistical significant differences on vocabulary gains across the groups using different output stimulus tasks?
3. Which type of tasks (e.g., fill-in-the-blank, sentence writing, composition, or a combination) most strongly affects L2 incidental vocabulary learning?
4. Does the amount of time spent on task have different effects on L2 incidental vocabulary learning?
5. Do the types of text (e.g., narrative, expository, or a combination of the two) have different effects on L2 incidental vocabulary learning?
6. Do text-target word ratios (≤2%, between 2% and 5%, and ≥5 %) have different effects on L2 incidental vocabulary learning?

Methodology

Dependent Variable

The dependent variable in this meta-analysis was L2 incidental vocabulary learning in an experimental-control group comparison design. The control group was asked to read a text without being asked to perform any additional tasks. On the other hand, the experimental groups were instructed to complete a task that required them to produce language output, such as fill-in-the-blank, sentence writing, composition writing, or a combination of these tasks. In addition to the task, some experimental groups were also given the same text to read as the control group, whereas some experimental groups were only asked to perform the output stimulus tasks without reading a text.

Sources of Data

The search for data sources had no beginning cut-off date; any research data available before the present study was considered a source. Studies investigating the effects of output tasks on L2 incidental vocabulary learning were collected through a variety of sources. Most were identified through computerized databases and search engines, including PsychINFO, ProQuest Dissertation and Theses, JSTOR, Education Resources Information Center (ERIC), Web of Science, Google Scholar, Linguistics+Language (CSA), Linguistics and Language Behavior Abstracts (LLBA), MLA Bibliography, Chinese Electronic Thesis & Dissertations Service (CETD), and Hong Kong University Theses Online. Terms, such as combinations of "incidental vocabulary learning," "foreign language," "second language," "reading," "output," "involvement load," and "task effectiveness" were used as search keywords for this meta-analysis study. Furthermore, citations in the reference sections of identified studies were consulted for cross-referencing to locate potential studies for inclusion. A total of 52 articles, book chapters, proceedings, dissertations, and theses published between 1975 and 2009 were retrieved based on these search parameters.

Inclusion and Exclusion Criteria

Each study met the following criteria for inclusion. First, they investigated L2—rather than L1—incidental vocabulary learning. Second, they used reading activities in which passages were used as reading materials. Third, they took the form of an experimental-control group comparison design, in which the effects of a task on incidental L2 vocabulary learning were measured. Fourth, this meta-analysis included

both published and unpublished studies, such as journal articles, book chapters, dissertations, theses, and technical reports. Most scholarly journals publish only research findings that are statistically significant. Omitting unpublished, statistically insignificant findings from this meta-analysis would therefore result in distorted, positive, high effects that may significantly skew the actual pattern of findings. Fifth, this current meta-analysis examined adult second and foreign language learners' L2 incidental vocabulary development. *Adult* is defined as a learner who is in a middle school or higher grade level.

On the other hand, studies that met the following criteria were excluded. First, studies were excluded if researchers provided a qualitative analysis without statistical reports or failed to report effect sizes or statistics for calculating effect sizes. Second, two studies, reprinted across several sources or based on the same sample, were used only once in this meta-analysis. For example, if a dissertation and a journal article were based on the same sample, the report, which contained more sufficient information, was included in this meta-analysis. Third, studies not written in English were excluded. Fourth, studies in which target words were explicitly taught to participants were excluded.

A total of 12 separate studies fulfilled these criteria. After retrieving and selecting these 12 studies, the researcher coded and classified the information they contained.

Behavioral science researchers generally collect data from individuals; the unit of analysis is an individual person (Lipsey & Wilson, 2001). However, in a meta-analysis,

the unit of analysis is an individual research study. If more than one effect size is generated from the same study (e.g., the same sample), these effect sizes are statistically dependent because they share historical and situational influences. Therefore, inclusion of more than one effect size per construct from a single study violates the assumption of independent data points; this will produce an inflated sample size and incorrect standard error estimates (Lipsey & Wilson, 2001). Thus, to follow the assumption of independent data points, only one effect size per construct from a single study was used in this meta-analysis to avoid an inflated sample size and inaccurate standard error estimates.

Coding Procedure

Each study was reviewed and coded for information required to calculate effect sizes and information related to the following five moderators: design quality, types of output task stimulus, time-on-task length, genres of text, and text-target word ratio. In addition, a coding sheet was developed for a detailed analysis of moderators for each study (See Appendix B).

Moderators Variables

Design Quality

Each study was coded for design quality, based on the Quality of Control Rating developed by Allen et al. (2009). Two components, quality of comparison group and statistical control, were considered while coding design quality.

Quality of the Comparison Group

For the quality of the comparison group (i.e., the output stimulus task group), three levels were coded to measure if the language proficiency level of the output

stimulus task groups matched with that of the non-output stimulus task group on one of the following levels: non-language proficiency level , mismatched language proficiency level, or equivalent language proficiency level. Level 1 compared the output stimulus task with non-output stimulus task group based on their non-language proficiency variables, such as years of learning English. At level 2, output stimulus and non-output stimulus task groups were not equivalent in their language proficiency levels; one group had higher language proficiency than the other. At level 3, these two groups had equivalent levels of language proficiency prior to the treatment. No significant differences in their language proficiency levels implied a great extent of overlapping of reasonable sample sizes in the distributions of these two groups (Allen et al., 2009), rather than implying identical means.

Quality of Statistical Control

A total of four levels of quality of statistical control were coded. A study that did not use a covariate was coded as level 1. At level 2, only distal covariates—non-vocabulary ability covariates such as years of learning a target language—were employed. At level 3, the covariate was a proximal measure of target words (e.g., scores on a vocabulary test), a measure different from the posttest. At level 4, the covariate was the pre-treatment measure of the outcome, which was the identical measure used in the repeated measure designs or in the general linear model.

Design quality was rated from 1 to 3 after the researchers jointly considered the rating for the comparison group and statistical control (See Table 1). Design quality was considered low if the study was coded as 1, and high if it was coded as 3. Design quality

was rated as 3 if a study used a repeated measure or used a general linear model to assess the pretest of target words and used a comparison group with an equivalent language proficiency level (quality of match=3) or with mismatched language proficiency (quality of match=2). The rationale was that the limitation of using a comparison group with mismatched language proficiency was partially compensated by the use of a comparison group with equivalent target word knowledge on the identical measure used in repeated measure designs. However, all else being equal, design quality was rated as 2 if the comparison group in the study matched the experimental group on non-language proficiency variables (quality of match=1). The rationale was that "the regression slope for the relationship between pretest and posttest within groups may violate the assumption of homogeneity" (Allen et al., 2009, p.488).

If the study did not use statistical controls or did not employ covariates of non-vocabulary ability (statistical control level 1 or 2), but used a comparison group with an equivalent language proficiency, the design quality was coded as 2, or the study design was rated as 1. This is because without the presence of a statistical control for language proficiency difference, using an equivalent comparison group to the control group on a measure of target words minimizes the probability that group differences of post-treatment are a result of third variables. These variables can affect both language proficiency and selection into the treatment intervention (Allen et al., 2009).

Table 1.
Determination of Quality of Control Rating

Statistical Control / Quality of Match	Covariate			Repeated Measure or GLM (4)
	None (1)	Distal (2)	Proximal (3)	
Matched on non-language-proficiency variables (1)	1	1	2	2
Mismatched language proficiency (2)	1	1	2	3
Matched on language proficiency level (3)	2	2	3	3

Note. Quality of Control—1: Low, 2: Medium, 3: High

Inter-rater Reliability

Two raters independently coded each study for design quality of each study. The percentage of agreement was .92 between raters. Disagreements with regard to coding between these raters were settled by re-examination of studies or discussion between the raters.

Types of Output Stimulus Tasks

Vocabulary exercises were used as moderators to represent the degree of output stimulus task load involvement. These exercises included fill-in-the-blank, sentence writing, composition writing, or a combination. Some studies used one type of exercise independently, while some combined two or more. Studies that used more than one type of output stimulus tasks were coded as a combination.

Time on Task

Different amounts of time on task for the output stimulus groups may have various effects on L2 incidental vocabulary learning. *Time on task* was defined as the amount of time that participants engaged in an assigned activity, such as fill-in-the-gap vocabulary exercises, sentence writing, composition, and/or reading a text. Based on the data reported in the studies, all other activities, such as the introduction of the study and the pretest/posttest of the target words were excluded from time on task.

Genres of Text

Each text that was used as a reading sample in a study was coded as a narrative, expository, or a combination of the two. If a study used both genres in its research design, it was coded as a combination.

Text-target Word Ratios

Each study was coded for its text-target word ratio, the average number of words of text surrounded by an individual target word. Laufer (1989) suggests that readers can approximately guess the meaning of an unfamiliar word if they have lexical knowledge of 95% of the text. According to Nation (2001), readers have to comprehend at least 98% of the text to accurately infer the meanings of unknown vocabulary. Thus, three ranges (≤2%, between 2% and 5%, and ≥5 %) were used to code the data regarding the text-target word ratio.

Data Analysis

Calculation of Effect Sizes

To address research questions about the effectiveness of task-involvement load on incidental vocabulary learning, the data from the 12 collected studies were calculated for Cohen's d (effect size), the confidence interval around the mean effect size, the homogeneity test of variance (*Q*-test), and fixed or regression analysis.

First, if studies did not provide effect sizes (Cohen's d, magnitude of an observed difference between two groups in standard deviations unit) of incidental vocabulary learning between the experimental and control groups, this was calculated by using means and standard deviation, *t*-test values, or *F*-test values. To determine the statistical significance of the aggregated effect size, a 95% confidence interval around the mean effect size was constructed for the immediate vocabulary test.

The homogeneity test of variance, or *Q*-test, was analyzed to determine if the variability among effect sizes was greater than the sampling error. A significant *Q*-test indicated that the variability among the effect sizes exceeded sampling errors and further analysis on moderator variables was warranted. Moderator variables (design quality, types of output task stimulus, time-on-task length, genres of text, and text-target word ratio) were further analyzed the variability among effect sizes.

Results

A total of 12 studies (six published and six unpublished articles) were reviewed based on the established coding sheet to construct a comprehensive view of the characteristics of these studies (See Appendix B). Of the 12 studies, nine (75%) had the

students' target language as English, and each of remaining three studies had either Spanish, German, or French as the target language. Table 2 summarizes the mean, standard deviation, effect size, and 95% confidence intervals around the mean effect size.

Inter-rater Reliability

To ensure the reliability of coding, the researcher coded information from each study. In addition, a fellow graduate student was trained and then independently coded the studies. The inter-coder reliability was .91. Disagreements in coding between the raters were resolved via re-examination of studies and discussion.

Design Quality Characteristics

Concerning the quality of comparison groups, eight studies were rated as low and four were coded as high. With regard to the quality of statistical control, 10 studies did not use research design with covariates (83.3%, coded as 1), and two studies used a repeated measure design with covariate (16.7%, coded as 4). Based on the joint consideration of quality of comparison group and statistical control, the design quality was determined and the results were as follows: eight studies coded as low, two studies as medium, and two studies as high. The mean quality of comparison group was 1.5 (SD=0.80). The rated design quality for each study is listed in Table 3.

Table 2.
The Mean and Standard Deviation, Effect Size, and 95% Confidence Intervals—Output Stimulus Tasks versus No Tasks

	Author (Year)	Output stimulus Group		Non-output stimulus Group		d	95% CI
		M	SD	M	SD		
1	Rott (2004)	3.00	0.71	1.82	0.60	1.78	0.02~1.81
2	Min (2008)	36.24	5.33	24.64	3.52	2.57	0.32~1.52
3	Hulstijn & Laufer (2001)	2.90	1.80	2.70	1.50	0.12	0.33~1.51
4	Keating (2008)	3.76	2.61	1.35	1.80	1.05	0.33~1.51
5	Kim (2008)	27.40	2.70	18.20	2.70	3.41	-0.02~1.86
6	Hulstijn & Trompetter (1998)	4.90	2.30	4.10	2.10	0.36	0.52~1.32
7	Brown (2003)	8.95	1.31	7.50	2.37	0.70	0.33~1.51
8	Lee (2002)	3.21	2.28	1.00	1.19	1.23	0.48~1.36
9	Beal (2007)	49.38	30.47	28.66	14.57	0.78	0.26~1.58
10	Tu (2004)	5.67	2.71	1.14	1.53	2.05	0.48~1.36
11	Hsu (2005)	177	32.02	144.44	32.06	1.01	0.32~1.52
12	Lan (2004)	14.72	3.22	10.93	3.95	1.05	0.37~1.47

Table 3.
Rated Design Quality— Output Stimulus Tasks versus No Tasks

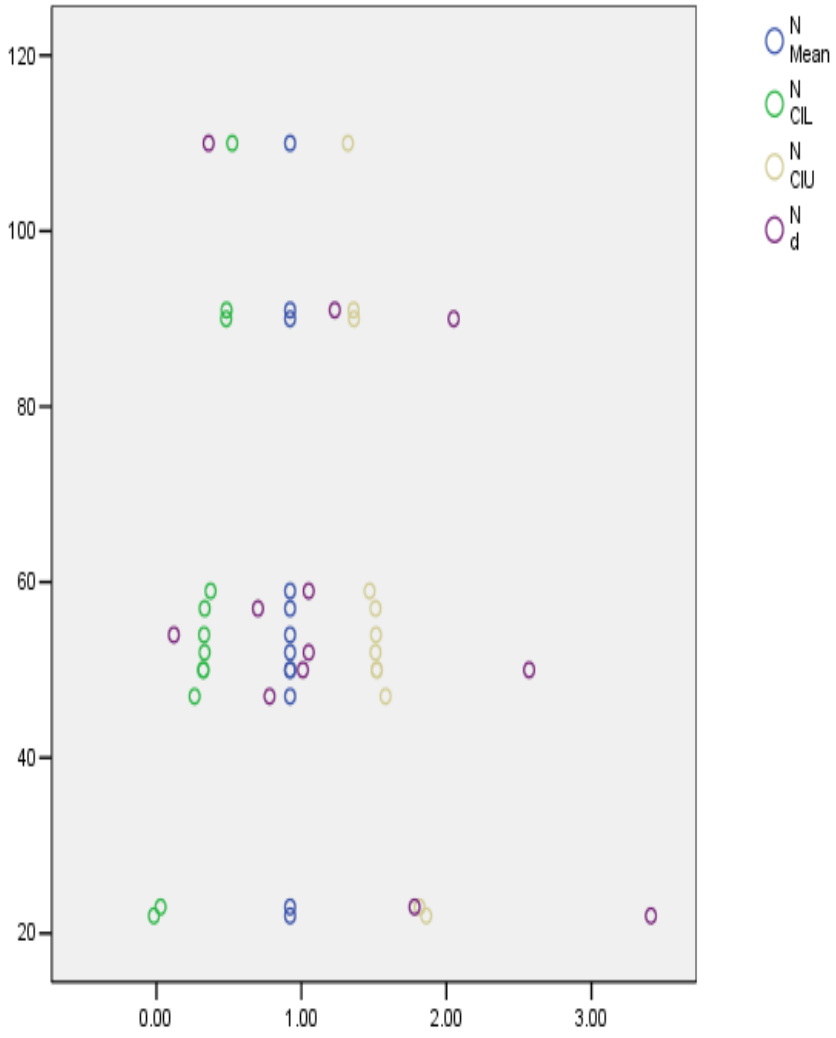
	Author (s)	# of control students	# of experimental students	Comp Group Quality	Stat Control Quality	Design Quality	Mean Hedges Effect Size
1	Rott (2004)	11	13	1	1	1	1.71
2	Min (2008)	25	25	3	4	3	2.53
3	Hulstijn& Laufer (2001)	20	33	1	1	1	0.12
4	Keating (2008)	23	29	1	1	1	1.04
5	Kim (2008)	33	14	1	1	1	3.28
6	Hulstijn& Trompetter (1998)	50	60	1	1	1	0.36
7	Brown (2003)	38	19	1	1	1	0.69
8	Lee (2002)	46	45	1	1	1	1.22
9	Beal (2007)	15	32	3	3	2	0.77
10	Tu (2004)	45	45	1	1	1	2.03
11	Hsu (2005)	25	25	3	4	3	1.00
12	Lan (2004)	87	80	3	1	2	1.04

Homogeneity of Effect Sizes

A test of homogeneity of variance (Q-test) indicated significantly more variation among effect sizes than one would expect from sampling error for incidental vocabulary learning (Q=73.6, df=11, p<.001). The average effect size was significantly different with confidence intervals ranging from 0.096 to 2.00. A funnel plot displaying sample size by effect sizes provided visual representations (see Figure 1). The significant result

detected in the *Q*-test warranted further investigation into the causes of effect size variations. Thus, the fixed-effect model was used to explore the causes of variations among effect sizes by investigating the moderating variables, whereas the random-effect model was employed to examine the causes.

Figure 1.
Funnel Plot— Output Stimulus Tasks versus No Tasks



Design Quality

Studies with high and medium levels of design quality were compared to those with a low level of design quality. Results revealed that a statistically significant difference in effect size magnitude was found between studies with high and medium levels of design quality and the low level (QB=76.33, df =1, p < .001). These results indicated that studies with high and medium design quality (M=2.29) were more likely to detect statistic significances on target-word learning than the studies with the low level of design quality (M=0.95).

Type of Tasks

A significant difference was found for the groups using different types of output stimulus tasks (e.g., fill-in-the-blank, sentence writing, composition, or a combination) to facilitate vocabulary learning (QB=11.53, df =3, p< .005). The groups of learners that used a combination of output stimulus tasks gained the highest mean effect size (M=1.60). The groups that used composition writing as the output stimulus task received the second vocabulary learning (M=1.06), followed by the groups that used sentence writing as the output stimulus tasks (M=0.94). The groups who used fill-in-the-blank exercises had the lowest mean effect size (M= 0.81).

Time on Task

Time on task ranged from 13 to 600 minutes. Hsu's (2007) study did not provide the amount of time used in the output task group and thus was excluded from the analysis. Results of the current study showed that time on task was significantly associated with vocabulary learning scores (B=4.29, p=0.001). This finding indicated

that language learners who spent more time engaging in the output task gained more vocabulary than those who spent less time on task.

Genres of Text

A significant difference in effect sizes based on genre of text was found (QB=12.00, df =2, p< .005). The mean effect size for the combination of various texts was the highest (M=1.01), followed by that for learners who read expository texts (M=0.7), and the narrative texts as the least effective (M=0.09).

Text-target Word Ratios

Two ranges of text-target word ratios (≤2% and between 2% and 5%) were identified among these studies; no study had a text-target word ratio greater than 5%. A fixed-effect model was used to explore whether vocabulary learning differed between those who read a text with text-target word ratios of ≤2% and those who read a text with ratios between 2% and 5%. The mean effect size for learners who read texts with text-target word ratios less than 2% (M= 1.3) was higher than for those who read texts with ratios of 2% to 5% (M=1.2). However, the difference was not statistically significant (QB=11.07, df =2, p< .005).

Discussion

This meta-analysis found that language learners who performed an output stimulus task scored greater vocabulary gains than those who did not. In addition, this study investigated the relationship between output stimulus tasks and moderator variables (design quality, types of output stimulus tasks, time-on-task length, genre of text, and text-target word ratio) of outcomes on incidental vocabulary learning. The

following moderators were found to be significant predictors of L2 incidental vocabulary learning: types of output stimulus tasks, time-on-task length, and genre of text.

The meta-analysis conducted by Wa-Mbaleka (2006) investigated eight different levels of depth of processing. No significant differences were observed in vocabulary gains among groups using various levels of depth of processing on L2 incidental vocabulary learning. Wa-Mbaleka (2006) attributed this result to the categorization of the depth of processing. Therefore, no comparison can be drawn between the results from Wa-Mbaleka (2006) and the current meta-analysis.

To answer research question 1 (RQ), results of the current meta-analysis showed that language learners who performed an output stimulus task gained more vocabulary than those who did not perform an output stimulus task. Results for RQ 3 further revealed that different gains in vocabulary were observed from learners who participated in different output stimulus tasks that required various involvement loads. Language learners engaged in a combination of output stimulus tasks outperformed those who did a task with a less extensive involvement load, such as fill-in-the-blank, sentence writing, or composition writing. These findings confirm the involvement load hypothesis that tasks with a higher degree of involvement load are more effective in L2 vocabulary learning. The findings also were consistent with previous studies (Beal, 2007; Hulstijn & Trompetter, 1998; Hulstijn & Laufer, 2001) showing that tasks with a more extensive involvement load better enhance L2 vocabulary learning than tasks with a less extensive involvement load.

Results for RQ 2 revealed that studies with medium and high levels of design quality— the integration of equivalent language proficiency levels and the repeated measures of target vocabulary — can better account for the variability in effect sizes as opposed to studies with a low level of design quality. Studies with medium and high levels of design quality had more statistical power to detect significances in the different magnitudes of effect sizes. Therefore, based on these results, future studies that use high and medium levels of design quality are warranted to better verify the research findings in the area of incidental vocabulary learning.

Results for RQ 4 showed that L2 incidental vocabulary learning corresponds to the amount of time spent on task. Time on task was positively correlated with L2 vocabulary learning. These findings supported those of Tu's (2004) study; language learners who spend more time on an output stimulus task gain more vocabulary. One possible explanation is that more time spent on task allows learners to connect newly encountered vocabulary more easily to their existing L2 lexical system.

Results for RQ5 showed that language learners who read a combination of expository and narrative texts gained more vocabulary than those who read either an expository or narrative text. Additionally, expository texts were more effective in enhancing vocabulary learning than narrative texts. One possible explanation is that the logical connections between words in expository texts help language learners better infer the meaning of unknown words. Pretorius (2006) suggested that expository text is structured as a logical "binary membership between two or more text units" (p.434). Logical relations can be causal links (e.g., *because*, *as a result of*, *consequently*, and

thus), positive propositions (e.g., *and*, *in addition*, *moreover*, and *furthermore*), negative propositions (e.g., *however* and *although*), illustrative relations (e.g., *for example* and *for instance*), or adversative relations (e.g., *yet*, *conversely*, *but*, and *on the other hand*). These logical, clear structures help learners derive the meaning of unknown words. However, nearly all of the studies that were included in the current meta-analysis used an expository text as the reading text. In this regard, these findings should be interpreted as tentative and with caution due to the small sample size for each type of reading text. Results for RQ 6 revealed that vocabulary gains did not differ significantly between language learners who read the texts with a ratio of ≤2% and those who read the texts with ratios between 2% and 5%. These results suggest that text-target word ratio did not have a significant effect on learners' vocabulary gains.

Conclusion

The results of this meta-analytic investigation clearly indicate that output stimulus tasks enhance second language vocabulary acquisition. An output stimulus task induces a higher level of involvement load than a reading task alone; language learners gained more vocabulary knowledge through engagement with output stimulus tasks. This supports the involvement load hypothesis. Among output stimulus tasks, learners who engaged in tasks with different levels of involvement load also showed different vocabulary gains. The combination of output stimulus task requires the strongest need, search, and evaluation, as opposed to fill-in-the-blank, sentence writing, or composition writing. In this regard, this task also yields the most demanding involvement load and requires learners to process lexical items at their deepest levels. As a result, learners who

engaged in this task gained the most vocabulary. Studies with different levels of design quality have been shown to have various effects in detecting significant differences in effect sizes; studies with medium and high levels of design quality can better detect the variation in effect sizes. Therefore, vocabulary researchers should use a research design with high and medium levels of design quality that include groups with equivalent language proficiency levels and use repeated measures of target vocabulary.

Based on the results of this study, time spent on task was also found to be an important factor in L2 vocabulary development. The more time language learners spent on task, the more vocabulary they acquired. Moreover, the type of text also had an impact on L2 incidental vocabulary learning. A combination of texts has the largest impact on L2 incidental vocabulary learning compared to an expository or narrative text alone. Additionally, the results showed that a combination of expository and narrative texts better facilitates L2 vocabulary learning than either an expository or narrative text alone. Finally, no difference was observed in vocabulary gains between language learners who read texts with the ratio of ≤2% and those who read texts with ratios of between 2% and 5%.

Limitation and Suggestions for Future Studies

Studies whose results are not statistically significant are more likely to be rejected by journals (Abraham, 2008; Norris & Ortega, 2006). A meta-analysis that only includes published studies with significant, high effect sizes may produce an inflated view of the results (Norris & Ortega, 2006). In this sense, an extensive search was conducted; attempts were made to include relevant unpublished studies in the current

investigation. Despite this effort, not all such unpublished research was included in this meta-analysis.

Results from this meta-analysis found that with the use of output stimulus tasks, language learners gained more vocabulary from reading a combination of expository and narrative texts than from either an expository or narrative text alone. However, the small sample sizes for each type of text (one for narrative text, one for a combination of these two types, and ten for expository text) make this result tentative rather than definitive. Significant differences in L2 vocabulary gains may exist between language learners who read expository and narrative texts but the small sample size for each type of reading texts may not have enough statistical power to detect the significance. Therefore, more studies on the effects of output stimulus tasks with different types of texts used as the reading material are needed. Additional studies will help clarify this research domain. Moreover, other moderators that may affect learners' L2 incidental vocabulary learning with the use of output stimulus tasks have not yet been adequately explored. Further research should investigate factors regarding individual differences (such as learners' second language proficiency level, vocabulary size, interest, and motivation) and how they are associated with L2 incidental vocabulary learning while performing output stimulus tasks. Additionally, further study should explore the role of different parts of speech, such as prefixes, suffixes, and roots, in L2 incidental vocabulary learning with the use of output stimulus task. Furthermore, future study should examine the long term effect of output stimulus tasks. Based on the results of the immediate vocabulary test, this study found that language learners' vocabulary was facilitated with the use of output

stimulus tasks; language learners doing output stimulus tasks outperformed those who had not engaged in output stimulus tasks. However, the long term effect of these tasks should be explored to determine whether or not the advantages of output use remain over time.

CHAPTER III

THE EFFECTS OF L1 AND IMAGE-BASED GLOSSES ON L2 INCIDENTAL VOCABULARY LEARNING WHILE READING — META-ANALYSES

Overview

The present study used meta-analytic techniques to investigate the effects of L1 textual and image-based glosses on second language (L2) incidental vocabulary learning while reading. Results showed that language learners gained more benefit from using textual glosses to learn vocabulary than those who had no access to glosses. Results indicated that text-target word ratios played a large role in second language vocabulary learning. Language learners who read a passage with a text-target word ratio of ≤2% gained more vocabulary than those who read a passage with a text-target word ratio between 2% and 5%. No statistically significant difference was detected between the groups that had access to multiple-choice and single glosses. Compared to paper-and-pencil environments, computer-assisted settings did not significantly facilitate L2 vocabulary learning. Language learners who were exposed to narrative reading materials did not significantly outperform those who read expository texts in relation to incidental vocabulary learning. No significant difference in L2 vocabulary learning was detected between groups who were provided with L1 textual glosses and those given L1 textual+image-based glosses.

Introduction

Vocabulary plays a critical role in second language learning. Nation (2006) suggests that a language learner needs a vocabulary of 8,000 to 9,000 word families to comprehend written text, such as a novel or a newspaper, and knowledge of approximately 6,000 to 7,000 word families to understand a typical movie. This requirement makes it impossible for learners to acquire language skills solely through explicit language learning. Thus, incidental vocabulary learning serves a complementary role in helping language learners increase their vocabulary, and extensive L2 reading serves a complementary role. Although inferring the meaning of an unknown word enhances vocabulary retention, learners are also likely to infer an inaccurate meaning. In addition, language learners may fail to connect the word form with its meaning in context. Coady (1993) suggests that vocabulary learning occurs when learners notice the meaning of a word and make a connection between the word form and its meaning. To minimize the chance of inferring an incorrect meaning of an unknown word, glosses are widely used in L2 vocabulary learning.

Literature Review

This section will provide an overview of the effects of textual and imaged-based glosses on L2 incidental vocabulary learning while reading. The overview is organized into three parts: the theoretical framework of textual and image-based glosses, previous studies on the effects of glosses on L2 incidental vocabulary learning, and an introduction of previous meta-analysis.

According to Nation (2001), textual glosses help language learners consolidate knowledge of unknown words. Textual glosses facilitate greater and more accurate comprehension of vocabulary (Jacobs, 1994). Nation (2001) proposes three advantages of textual glosses in second language learners' vocabulary development. First, language learners with access to glosses can increase their reading comprehension of text that otherwise is beyond their proficiency level. Second, glosses draw readers' attention to unknown words. Third, glosses can help readers learn independently and rely less on teachers for explanations. However, guessing strategies are considered vital for developing good reading comprehension. Although conventional single textual glosses can foster vocabulary learning, the meaning is given, which discourages learners from guessing the meaning of an unknown word based on its context. Research (Hulstijn, 1992; Hulstijn & Laufer, 2001) has shown that learners retain more vocabulary when they must infer the meaning of a word. This may be due to the fact that the process of inferring requires mental effort, which is positively correlated with learners' information recall and retention (Hulstijn, 1992; Hulstijn & Laufer, 2001). The more mental effort that is expended, the more new vocabulary learners retain. However, learners with access to single glosses do not use as much mental effort in reading as those without such access. To address this issue, Hulstijn (1992) proposed the use of multiple-choice glosses, which combine the advantages of inference with single glosses because they reduce the challenges imposed by insufficient context and the possibility of inaccurate inferences.

In addition to textual glosses, image-based glosses are widely used to enhance L2 incidental vocabulary learning. Images provide links between words and concepts and provide a connection between an unfamiliar word and its concept. Dual Coding Theory proposed by Paivio (1971) investigates how information is represented in the mind. Information is stored as images and verbal representations in two distinctive yet complementary subsystems. The non-verbal system is referred to as the imagery system (Paivio, 1971), which involves the representation and processes pertaining to non-verbal objects. The imagery system analyzes scenes, yields images, and involves sensory modalities such as environmental sounds and textures of objects. In contrast, the verbal system deals with representations and processes concerning language, such as auditory and visual representations of words. Dual Coding Theory suggests that language learning is promoted when learners use materials that feature both the imagery and verbal systems.

Previous Research on Glosses

Many studies have investigated the effects of glosses on L2 incidental vocabulary learning under different premises, such as the overall effects of textual, L1 or L2 textual, single and/or multiple-choice glosses, and imaged-based glosses. This section provides an overview of the studies that explored the effects of the previously mentioned types of glosses on L2 incidental vocabulary learning.

The overall effects of glosses have been widely tested under different conditions, including the effects of L1 or L2 textual, and single and/or multiple-choice textual or image-based glosses. Empirical evidence shows that glosses positively influence L2

incidental vocabulary learning. Research (Hulstijn, 1992; Hulstijn et al., 1996; Ko, 1995; Rott, 2004; Watanabe, 1997) has shown that textual glosses facilitate second language learners' vocabulary growth. Language learners with access to glosses, regardless of the type, outperform those who are not provided any glosses.

Beginning learners generally prefer glosses in their native language rather than the L2 (Taylor, 2006). Nation (2001) suggests that both L1 and L2 glosses are effective in facilitating vocabulary learning if the glosses are easily comprehended. Yoshhii (2006) found that both L1 and L2 glosses are beneficial for incidental vocabulary learning. However, language learners with access to L1 textual glosses did not perform significantly better than those who were provided L2 glosses. Other research (Krantz, 1991; Oskarasson, 1975; Scherfer, 1993) indicated that compared to L2 glosses, L1 glosses are more effective in second language vocabulary learning.

Some research has compared the effects of providing language learners with single textual or multiple-choice glosses (Hulstijn, 1992; Rott, 2004; Watanabe, 1997). Hulstijn (1992) conducted several studies to compare the impact of single textual glosses or multiple-choice glosses on L2 incidental vocabulary learning. Hulstijn (1992) found that language learners who have to infer the meaning of unknown words in text by themselves (high mental effort) are more likely to remember the form and meaning of an unfamiliar word compared to those who are given the meaning of unknown words. In addition, language learners are more likely to infer an inaccurate meaning of an unknown L2 word in a text when no cues pertaining to the meaning are provided.

However, other studies yielded different results (Wang, 2005, Watanabe, 1997). Wang (2005) investigated the effects of single glosses and multiple-choice glosses on L2 incidental vocabulary learning and found that groups provided single glosses and multiple-choice glosses performed significantly better on a word recall posttest than the group that received no textual glosses. The group using single glosses outperformed the group using multiple-choice glosses in the first posttest (seven days after the study) but not in the delayed posttest (14 days after the study). Learning vocabulary incidentally with single glosses thus appears to be more effective than multiple-choice glosses in the short term, but this effect diminishes over the longer term. Similarly, Watanabe (1997) found that groups that received single glosses and multiple-choice glosses outperformed the group without access to any glosses. In addition, no statistically significant difference was detected between groups using single glosses and multiple-choice glosses.

In summary, textual glosses are effective in L2 vocabulary learning. However, no consensus has been reached with regard to the effects of single glosses and multiple-choice glosses. Some studies indicated that language learners with access to single glosses demonstrated greater vocabulary growth than those provided multiple-choice glosses, whereas other research yielded different results. Hulstijn (1992) suggested that the discussion regarding glosses in L2 vocabulary acquisition should focus on which kind of glosses are most effective (a single definition in L1 , synonym in L2, multiple-choice definitions in L1 or L2, or a combination of different kinds of glosses), rather than on whether glosses should be provided at all.

Technology has changed our culture of reading and has added a new dimension to second language acquisition. As Gambrell (2005) pointed out, "the computer has put information literally at our fingertips" (p.589). Similarly, on-screen reading has become increasingly important. Legibility, or "the adequacy between a given text and its intended readership" (Macedo-Rouet et al., 2003, p.101), varies with different presentation media, such as a computer screen or paper. Legibility can be categorized as either surface or deep. According to Labasse (1999), surface legibility is associated with the perceptual distinctions of a text, such as font size, color, and contrast, while deep legibility concerns the semantic and structural factors that determine a text's level of comprehensibility. Texts presented in hypermedia, or on computers, are sometimes nonlinear in that they can be browsed in multiple configurations (Wilhelm, 2000). Books generally are read sequentially, from page 1 to page 2 and so on. Reading from a screen is more likely to be nonlinear in that readers can click on a link to access image-based glosses of an unknown word, then access pictorial glosses via a separate window. Zumbach and Mohraz (2008) found that learners considered reading a narrative text via hypermedia much more challenging than reading it through a linear format. The nonlinearity of on-screen reading sometimes can result in reader disorientation, high cognitive load, and low reading comprehension (Dillon & Gabbard 1998). In addition, reading from a computer screen may retard reading speed (Kerr & Symons, 2006) or disrupt the reader's mental maps of the text, which might lead to poorer understanding or recall of the material (Kerr & Symon, 2006).

However, Pomplun et al. (2002) argued that this slower rate was related to technological variables that have improved in recent years, such as font sizes and screen resolutions. After comparing the effects of reading comprehension using hypermedia and paper, Dillon and Gabbard (1998) found no significant differences between presentation media. Zumbach and Mohraz (2008) found that learners did not gain significantly different knowledge acquisition or had different levels of cognitive load while reading an encyclopedia text through different presentation formats (linear or non-linear).

Others (Bowles, 2004; Chun & Plass, 1996; Al-Seghayer, 2001; Yoshii, 2006) have investigated the effect of textual glosses on vocabulary learning for second-language learners in computer-assisted environments. Some studies (Chun & Plass, 1996; Al-Seghayer, 2001; Yoshii, 2006) found that multimedia glosses with vocabulary definitions facilitated language learners' vocabulary development. In contrast, Bowles (2004) found no significant results between learners' vocabulary gains using L1 glosses in computer-assisted versus paper-and-pencil environments.

The growing availability of computers has diversified the use of glosses. As a result, visual representations (namely, image-based glosses) have been incorporated in L2 vocabulary learning. The study of image-based glosses likewise has increased. Several researchers (Al-Seghayer, 2001; Chun & Plass, 1996; Kost, Foss, & Lenzini, 1999; Kuo & Chiang, 2006; Yanguas, 2009; Yoshii, 2006) have investigated the combined effect of L1 textual and image-based glosses on L2 vocabulary learning. Chun and Plass (1996) found that language learners exposed to L1 textual+pictorial glosses

demonstrated greater L2 vocabulary gains than those exposed only to L1 textual glosses or to L1 textual+video glosses.

On the other hand, Al-Seghayer (2001) concluded that L2 language learners showed greater vocabulary gain if the words were provided with L1 glosses+video clips rather than glossed only with pictures. In addition, learners presented with text+video glosses outperformed those presented with the text+picture. Similarly, Kuo and Chiang (2006) investigated the effects of three types of multimedia glosses—textual, pictorial, and textual+animation—on L2 incidental vocabulary learning. Participants were randomly assigned into one of four groups: control, textual, animations, and textual+animation. Each participant was asked to read a text for comprehension. In addition, participants who were not in the control group received various types of glosses for target words. Those in the textual glosses group received textual glosses, those in the pictorial group received animated glosses, and those in the textual+animation group received both textual and animated glosses. Immediately after reading the text, participants were instructed to complete an unexpected vocabulary test. Participants in the textual+animation group performed the best, followed by the textual, pictorial, and control groups.

Yoshii (2006) found that language learners who had access to L2 textual+pictorial glosses benefited more from the additional pictures than those provided with L1+pictorial glosses, suggesting that cues from pictorial glosses better delineate the meaning of a word. The L1 textual cues were sufficiently clear and additional pictures that depicted the same information were unnecessary. Therefore, language learners did

not increase their understanding after seeing additional pictures while L1 textual glosses were available to them. On the other hand, L2 textual information was not processed as well as L1 glosses. Thus, pictures supplemented or strengthened learning when L2 textual glosses were used.

In a more recent study, Yanguas (2009) also investigated the effect of multimedia glosses on L2 incidental vocabulary learning. Two weeks prior to the experiment, all participants took a vocabulary pre-test. Participants were randomly assigned to one of four groups: control, textual, pictorial, and textual+pictorial glosses. Results showed that all groups exposed to glosses outperformed the control group in the immediate vocabulary recognition posttest. In addition, participants who were provided with textual+pictorial glosses performed the best, followed by the pictorial glosses, textual glosses, and control groups.

In summary, imaged-based glosses are beneficial to L2 incidental vocabulary learning. However, a comprehensive picture comparing the effect of L1 textual and L1 textual+image-based glosses, which the experiments failed to provide, is warranted. As Yoshii (2006) suggested, ascertaining whether the combined effect of textual and visual glosses is redundant or supplemental may enhance our ability to facilitate L2 vocabulary learning.

Previous Meta-analysis

A meta-analysis provides a systematic approach to characterize patterns in quantitative studies, but patterns can be difficult to detect in studies that use a narrative approach. Abraham's (2008) meta-analysis examined the effect of computer-mediated

glosses on L2 incidental vocabulary learning by comparing six studies. Abraham (2008) reported that computer-mediated glosses have an overall effect size of 1.40 for the immediate vocabulary posttest, whereas these studies for the delayed vocabulary posttest produced an overall effect size of 1.25. Regarding the effect of glosses on learners' proficiency levels, those at the intermediate proficiency level benefit the most and beginning learners the least. However, the difference between levels of instruction was not statistically different.

Although this meta-analysis shed light on the effects of L1 glosses on L2 incidental vocabulary learning, some questions remain. The meta-analysis focused only on studies conducted in computer-assisted environments. The magnitude of the effects of L1 glosses on L2 vocabulary learning conducted in a paper-and-pencil setting might be different. In addition, many moderators to vocabulary learning, such as the ratio of text-target words and multiple-choice glosses, were not investigated. Moreover, the effects of image-based glosses on incidental vocabulary learning have not been investigated. Building on previous work, the current investigation was designed to provide more comprehensive meta-analytic reviews of the effects of L1 and image-based glosses on L2 incidental vocabulary learning with a focus on receptive vocabulary.

Research Questions

The current study attempted to address the following research questions:

1. What are the overall effects of L1 textual glosses and image-based glosses on L2 incidental vocabulary learning?

2. Does the design quality (low, medium, or high) of the study have different effects of detecting statistically significant differences among the groups with different glosses (L1 textual glosses, L1 textual+image-based glosses, or no glosses)?
3. Do the uses of multiple-choice glosses and single glosses in reading cause varying effects on L2 incidental vocabulary learning?
4. Does the type of setting used (computer-assisted or paper-and-pencil) have varying effects for different types of glosses on L2 incidental vocabulary learning?
5. Do the text-target word ratios (≤2% and 2% ~ 5%) have varying effects on L2 incidental vocabulary learning based on different types of glosses?
6. Does the genre of reading text (narrative and expository texts) have different effects based on the types of glosses on L2 incidental vocabulary learning?

Two meta-analyses were conducted to answer these research questions. In the first, the experimental group had access to L1 textual glosses, whereas the control group had no access. In the second, the experimental group had access to L1 textual+image-based glosses and the control group had access to L1 textual glosses.

Methodology

Dependent Variable

The dependent variable in these meta-analyses was L2 incidental vocabulary learning in an experimental-control comparison. In the first meta-analysis, the control group was the group that had no access to any types of glosses, whereas the experimental group was the group that was provided with definitions of the target words in participants' first language. In the second meta-analysis, the experimental group was

given the textual+pictorial glosses, whereas the control group had access only to the L1 textual glosses.

Sources of Data

In the present investigation, many studies were identified via computerized databases and search engines, including PsychINFO, ProQuest Dissertation and Theses, JSTOR, Education Resources Information Center (ERIC), Web of Science, Google Scholar, Linguistics+Language (CSA), Linguistics and Language Behavior Abstracts (LLBA), MLA Bibliography, Chinese Electronic Thesis & Dissertations Service (CETD), and Hong Kong University Theses Online. The following keywords were used as search terms: gloss, glosses, glossary, annotation, dictionary, electronic dictionary, imaged-based glosses, incidental vocabulary learning, second language learning, vocabulary development, and foreign language vocabulary. In addition, many studies (Ben Salem, 2006; Bowles, 2004; De Ridder, 2003) were identified from citations in the previous meta-analysis (Abraham, 2008). Some of those journals could not be located using electronic databases. The table of contents from relevant journals, such as *Foreign Language Annals* and *Studies in Second Language Acquisition*, were manually reviewed. Cross-referencing references cited in literature yielded a few studies (Ko, 1995; Kuo & Chiang, 2006). A total of 98 articles, book chapters, proceedings, dissertations, and theses were retrieved.

Inclusion and Exclusion Criteria

Eligible studies in these meta-analyses satisfied three criteria. First, they focused on L2 vocabulary learning. Studies investigating the effect of L1 vocabulary learning

were excluded. Second, they used passage reading; sentence-level reading was beyond the scope of this investigation. Third, both published and unpublished studies were considered eligible, including journal articles, book chapters, dissertations, theses, and conference proceedings. Scholarly journals typically publish research results that are statistically significant with high, positive effect sizes. However, if a meta-analysis only included published articles with significant, high effect sizes, an inflated view will occur, resulting in a considerable skew in the research domain. Therefore, the current investigations included both published and unpublished studies.

The current meta-analyses focused solely on learning vocabulary while reading. Some studies were excluded. First, studies with a research design that did not include reading tasks were excluded. Second, studies written in a language other than English were excluded. Third, studies in which target words were introduced and/or taught explicitly to either the control or experimental group were not included. Fourth, studies were excluded if the authors did not provide sufficient statistical information for computing effect sizes. Fifth, studies in which the research design did not meet the research topic in this investigation were also excluded. For example, Holley and King's (1971) study was excluded because it investigated the effect of the location for glosses on L2 vocabulary learning. In addition, Al-Seghayer's (2001) study in which participants were exposed to different types of glosses for different words was excluded. Finally, the dataset of a study reprinted in more than one paper was only used once.

Two meta-analyses were included in the current meta-analysis study: 16 studies that fulfilled the criteria for investigating the effects between the textual glosses and no

glosses groups were used for the first meta-analysis, and eight studies that met the criteria for the effects between the L1 textual+image-based glosses and L1 textual glosses groups were used for the second meta-analysis.

Coding Procedure

Each study was coded for information required to calculate effect sizes and information concerning the following moderator variables: quality of the study design, computer-assisted/paper-and-pencil-based environment, text-target word ratio, multiple-choice/single glosses, and genre of reading text (see Appendix C).

Moderator Variables

Design Quality

Each study was coded for design quality based on the quality of control rating developed by Allen et al. (2009). Two components were considered: quality of comparison group and statistical control. For the quality of the comparison group, three levels (non-language proficiency level, mismatched language proficiency level, or equivalent language proficiency level) were coded to assess whether the language proficiency level of the experimental and control groups were equivalent (see Table 4 on page 60). At Level 1, learners from the comparison group and control group matched on non-language proficiency variables, such as majors and years of receiving target language instruction. At Level 2, learners from the comparison group and the control group did not have an equivalent language proficiency level; learners from one of the groups had a lower language proficiency level. Finally, at Level 3, no statistically significant differences were observed in these two groups of learners' language

proficiency level prior to the treatment. It should be noted that equivalent language proficiency levels between these groups does not imply identical means, but rather that a reasonable amount of overlap exists in the distributions of these two groups (Allen et al., 2009).

Quality of Statistical Control

Four levels of statistical control were coded. At the lowest level, no covariates were used. At Level 2, distal covariates, defined as non-language proficiency covariates, were employed. Distal covariates included years of learning English or average age of participants. At Level 3, proximal covariates, termed as a measure of target words (such as scores on a vocabulary test), were employed. At Level 4, the covariate was the pre-treatment measure of the outcome as in repeated measures or in general linear models.

Three levels of design quality (low, medium, and high) were rated after considering both ratings for the quality of comparison group and statistical control (see Table 4). If design quality for a study was rated as 1, the study was considered of low design quality, whereas if the design quality was coded as 3, the study had a high degree of design quality. As Allen et al. (2009) noted, "The reason for considering these two strategies for removing selection differences between the experimental and control groups is that neither strategy is a perfect remedy to non-random assignment" (p.488). For instance, if a study had a statistical control for the pretest on target word knowledge as in repeated measures (statistical control level=4), and the study employed a comparison group with an equivalent or mismatched language proficiency level (quality of match=2 or 3), the study was rated as high quality. The use of a comparison group

with equivalent target word knowledge on the same measure used in repeated measures partially compensated for the limitation of using a comparison group with mismatched language proficiency. With all else being equal, if its comparison group matched the non-language proficiency level with the control group (quality of match=1), the design quality was coded as 2 due to concern that "the regression slope for the relationship between pretest and posttest within groups may violate the assumption of homogeneity" (Allen et al. 2009, p.488).

If the study did not employ the same vocabulary test in the pretest as the posttest (statistical control level=3) but used a comparison group with an equivalent language proficiency level as the control group (quality of match=3), design quality was rated as 3. This is because the use of a comparison group with an equivalent language proficiency level partially compensates for the limitations of using a measure of target word knowledge different from the outcome measure (Allen et al., 2009). However, with all else being equal, if a comparison group with a lower language proficiency level was used (quality of match=1 or 2), the design quality was coded as 2.

If the study failed to use statistical controls or covariates as non-target word knowledge variables (statistical control level 1 or 2), but the comparison and control group had an equivalent language proficiency level, the design quality was rated as 2. Otherwise, the study was rated as1. According to Allen et al. (2009), without the presence of a statistical control for target word knowledge differences, employing an equivalent comparison group of language proficiency "minimizes but does not eliminate the probability that group differences in the post-treatment are a result of third variables"

(p.489). The third variables may influence both vocabulary learning and selection into the intervention.

Table 4.
Determination of Quality of Control Rating

Statistical Control / Quality of Match	Covariate			Repeated Measure or GLM (4)
	None (1)	Distal (2)	Proximal (3)	
Matched on non-language-proficiency variables (1)	1	1	2	2
Mismatched language proficiency (2)	1	1	2	3
Matched on language proficiency level (3)	2	2	3	3

Note. Quality of Control=1: Low, 2: Medium, 3: High

Single/ Multiple-choice Textual Glosses

Different textual glosses may have different effects on language learners' vocabulary development. Single glosses and multiple-choice glosses are commonly used to facilitate L2 vocabulary learning. Therefore, in the current meta-analyses, each study was coded to identify it as using single or multiple-choice textual glosses.

Computer-assisted versus Pencil-based Environment

Technological advances have greatly changed second language learning. Computers have increasingly assisted with L2 vocabulary learning. To investigate the effect of presentation medium on L2 vocabulary learning, each study was coded (1 or 2) to identify it as a computer-assisted or pencil-and-paper based environment.

Text-Target Word Ratios

The text-target word ratio is the average number of words of text accompanied by individual target words. Three ranges of text-target word ratios (≤2%, between 2% and 5%, and ≥5 %) were investigated in the current meta-analyses. Laufer (1989) suggested that adequate reading comprehension occurs when language learners comprehend at least 95% of the words in a text, allowing them to approximately guess the meanings of unknown or unfamiliar words. Nation (2001) recommended that readers should be familiar with 98% of the words in a text. Therefore, each study was coded for its text-target word ratio and categorized into one of three ranges (≤2%, between 2% and 5%, and ≥5 %). These ranges were investigated for the effect of glosses on L2 vocabulary learning.

Genres of Text

Genre of reading text (narrative or expository) may have different effects on language learners' vocabulary development. Vocabulary in narrative texts may be easier to learn than in expository texts, because words in narrative texts are more concrete and less content-specific. In addition, narrative texts have less information density in terms of parts of speech. Therefore, language learners are more likely to learn words in narrative texts. Thus, the reading text used in each study was coded as expository or narrative.

Data Analysis

Inter-rater Reliability

The researcher first coded the data based on the quality of control rating (see Table 4) and the coding sheet for each study (see Appendix C). To estimate reliability of coding, a fellow graduate student, who was trained by the researcher, independently coded studies that were included in the current meta-analyses. The percentages of agreement between these two coders were .94 for the rating of design quality and .91 for the other moderator variables. Disagreements in coding between the raters were settled via re-examination of the studies and discussion of the discrepancies.

Calculation of Effect Sizes

The effect size measure used in the current meta-analyses, Cohen's d, was the mean group difference, which is the mean immediate vocabulary posttest score of the experimental group subtracted from that of the control group, divided by the pooled standard deviation. Hedge's correction was also calculated. All effect sizes were weighted by the inverse variance and averaged to create the overall effect size. To determine the statistical significance of the effect size, a 95% confidence interval was constructed around each weight mean effect size for the outcome variable.

The homogeneous of variance test (Q-test) was calculated to determine whether the distribution of weighted effect sizes was normal and the results of sampling errors. If the result of homogeneous of variance test (Q-test) was statistically significant, the fixed effects model, analog to the Analysis of Variance, was used for categorical independent variables to further analyze the moderating variables.

Results

As already noted, two separate meta-analyses were conducted: the effects of L1 glosses on incidental vocabulary learning, compared to the group without the provision of any glosses on incidental vocabulary learning, and differences between the effect of the group with L1 and image-based glosses and the group with L1 glosses on incidental vocabulary learning.

1st Meta-analysis: L1 Glosses on L2 Vocabulary Learning

In the meta-analysis on the effect of L1 glosses on L2 incidental vocabulary learning, the experimental group was the one with access to L1 glosses, whereas the control group had no access. Of the 16 studies, half used English as participants' target language, followed by Spanish (five), French (two) and German (one). For these 16 studies, the mean effect size was 1.39 (se=.07). Effect sizes ranged from 0.43 to 2.77 (see Table 5).

Table 5.
The Mean and Standard Deviation, Effect Size, and 95%—L1 Textual Glosses versus No Glosses

	Author (Year)	No Glosses Group		L1 Textual Gloss Group		d	95% CI
		M	SD	M	SD		
1	De Ridder (2003)	0.17	0.11	0.35	0.13	1.49	0.79~1.99
2	Jacobs et al. (1994)	5.3	2.0	9.6	4.1	1.29	0.81~1.97
3	Bowles (2004)	N/A		N/A		1.64	0.63~2.14
4	Watanabe (1997)	6.67	4.69	18.45	9.07	1.61	0.93~1.85
5	Ben Salem (2006)	0.50	0.71	2.53	1.68	1.56	0.65~2.13
6	Ko(1995)	14.1	2.59	22.13	3.18	2.77	0.82~1.96
7	Plass et al. (2003)	20.5	3.70	24.6	3.30	1.17	0.88~1.90
8	Wu (2002)	3.77	2.32	6.54	2.30	1.20	0.91~1.87
9	Huang (2003)	3.83	3.58	8.02	3.74	1.14	0.93~1.85
10	Hulstijn et al. (1996)	0.6	0.6	2.60	1.90	1.42	0.78~2.00
11	Chen (2004)	0.4	0.62	1.83	1.62	1.17	0.81~1.97
12	Knight (1992)	8.75	3.38	14.56	4.49	1.46	0.96~1.82
13	Beal(2007)	28.66	14.57	58.61	27.37	1.23	0.71~2.07
14	Chan (1999)	4.9	1.86	7.29	1.38	1.45	0.70~2.08
15	Yanguas (2009)	8.17	1.55	13.55	3.03	2.28	0.70~2.08
16	Wang (2005)	1.58	1.96	2.48	2.26	0.43	0.68~2.10

Regarding the quality of the comparison group, 13 studies were rated as low and three studies were coded as high. Concerning the quality of statistical control, effects were computed using no covariates in 12 studies (coded as 1), proximal measure in one study, and repeated measures in three studies (coded as 4). Considering both the quality of the comparison group and statistical control, the design quality for each study was rated and the results were as follows: 10 studies were coded as low, five studies as medium, and the remaining study as high. The mean level of design quality across the studies was 1.50 (SD=0.73). The rated design quality for each study is listed in Table 6.

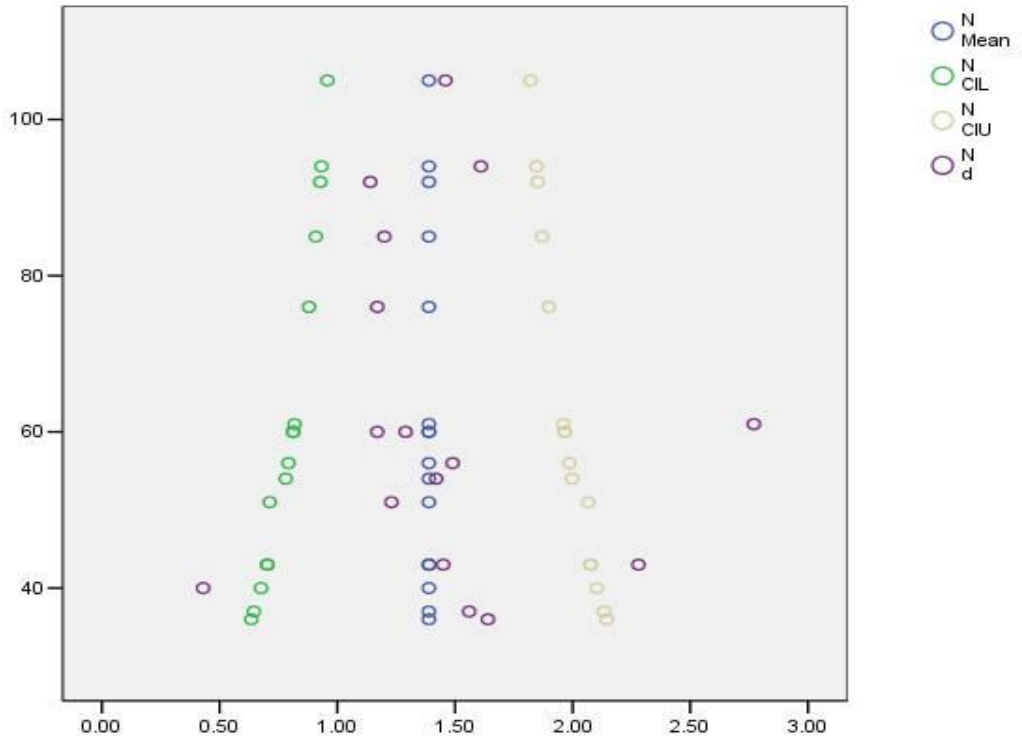
The test of homogeneity of variance (Q-test) indicated a significantly greater variation among effect size than one would expect for sampling errors for L2 incidental vocabulary learning (Q=34.48, df=15, p<.05). A funnel plot of this analysis also suggested that variations were more than sampling errors (see Figure 2). The result of heterogeneity in Q-test warranted further investigation of moderator variables.

Table 6.
Rated Design Quality— L1 Glosses versus No Glosses

	Author (s)	# of Control Students	# of Experimental Students	Comp Group Quality	Stat Control Quality	Design Quality	Mean Hedges Effect Size
1	De Ridder (2003)	28	28	1	1	1	1.47
2	Jacobs et al. (1994)	27	33	1	1	1	1.28
3	Bowles (2004)	18	18	1	1	1	1.60
4	Watanabe (1997)	45	49	1	4	2	1.60
5	Ben Salem (2006)	18	19	1	1	1	1.52
6	Ko(1995)	31	30	1	4	2	2.74
7	Plass et al. (2003)	38	38	1	1	1	1.16
8	Wu (2002)	44	41	3	1	2	1.19
9	Huang (2003)	46	46	1	1	1	1.10
10	Hulstijn et al. (2006)	27	27	1	1	1	1.40
11	Chen (2004)	30	30	3	1	2	1.15
12	Knight (1992)	51	54	1	1	1	1.45
13	Beal(2007)	15	36	3	3	3	1.21
14	Chan (1999)	22	21	1	1	1	1.43
15	Yanguas (2009)	23	20	1	4	2	2.25
16	Wang (2005)	20	20	1	1	1	0.42

Note. Comp Group Quality = Comparison group quality; Stat Control Quality = Statistical control quality

Figure 2.
Funnel Plot— L1 Textual Glosses versus No Glosses



ANOVA analyses were conducted to detect the difference among the studies with regard to moderator variables. Results showed that a statistical difference (QB=22.27, df =2, p<.05) was found among effect sizes for studies with various levels of design qualities. Studies with the highest level of design quality (M=2.74) could best detect significances in variations in effect sizes of L1 textual glosses on L2 vocabulary learning, followed by studies of a medium level of design quality (M=1.40) and studies of the lowest level of design quality (M=1.27).

No statistically significant difference (QB=2.15, df =1, p=0.14) was found between the groups that used multiple-choice (M=1.19) and single glosses (M=1.42). In

addition, no statistically significant difference (*QB*=0.52, *df* =1, *p=.47*) was detected between the groups that used computer-assisted settings (M=1.42) and paper-and-pencil environments (M=1.34). Two ranges of text-target word ratios were identified: (≤2% and between 2% and 5%). A significant difference (*QB*=4.46, *df*=1, p=0.03) was found among the groups that used text-target word ratios that were ≤2% (M=1.54) and between 2% and 5% (M=1.27). No statistical difference (*QB*=0.72; *df*=1, *p*<0.39) was detected between groups that used expository (*M*=1.44) and narrative (*M*=1.33) reading materials.

2nd Meta-analysis: L1 and Image-based Glosses versus L1 Glosses

Eight studies with 524 participants were included in the second meta-analysis. Of these, the participants' target language was English in five, German in two, and Spanish in one. The effect sizes were fairly small, ranging from 0.12 to 0.94 (See Table 7). The mean effect size was 0.38 (*Se*=0.08).

When the quality of the comparison group was analyzed, seven studies were rated as low (coded as 1) and the remaining study was coded as high (coded as 3). In the analysis of the quality of statistical control, effects were computed using no covariates in three studies (coded as 1) and five studies used covariates in repeated measures (coded as 4). After considering the joint quality of the comparison group and statistical control, the design quality of each study was rated: two studies were coded as low and six as medium. Table 8 lists the characteristics of the design quality for each study.

Table 7.
The Mean and Standard Deviation, Effect Size, and 95% Confidence Intervals —L1+ Image-based Glosses versus L1 Glosses

	Author (Year)	L1 Textual Glosses Group		L1 Textual+ Image-Based Glosses Group		d	95% CI
		M	SD	M	SD		
1	Plass et al.(2003)	24.6	3.30	25.4	4.90	0.19	-0.08~0.84
2	Wu (2002)	6.54	2.30	6.82	2.30	0.12	-0.06~0.82
3	Akbulut (2007)	26.78	6.45	28.91	4.00	0.40	-0.22~0.98
4	Hsu (2008)	14.00	3.89	16.29	3.20	0.64	-0.24~1.00
5	Yoshii & Flaitz (2002)	6.12	3.05	7.58	2.60	0.52	-0.02~0.78
6	Yoshii (2006)	7.87	2.78	8.54	3.14	0.23	-0.03~0.79
7	Yanguas (2009)	13.55	3.03	15.57	3.77	0.58	-0.23~0.99
8	Kost et al. (1999)	8.61	3.78	11.53	2.18	0.94	-0.31~1.08

Note: d=effect size

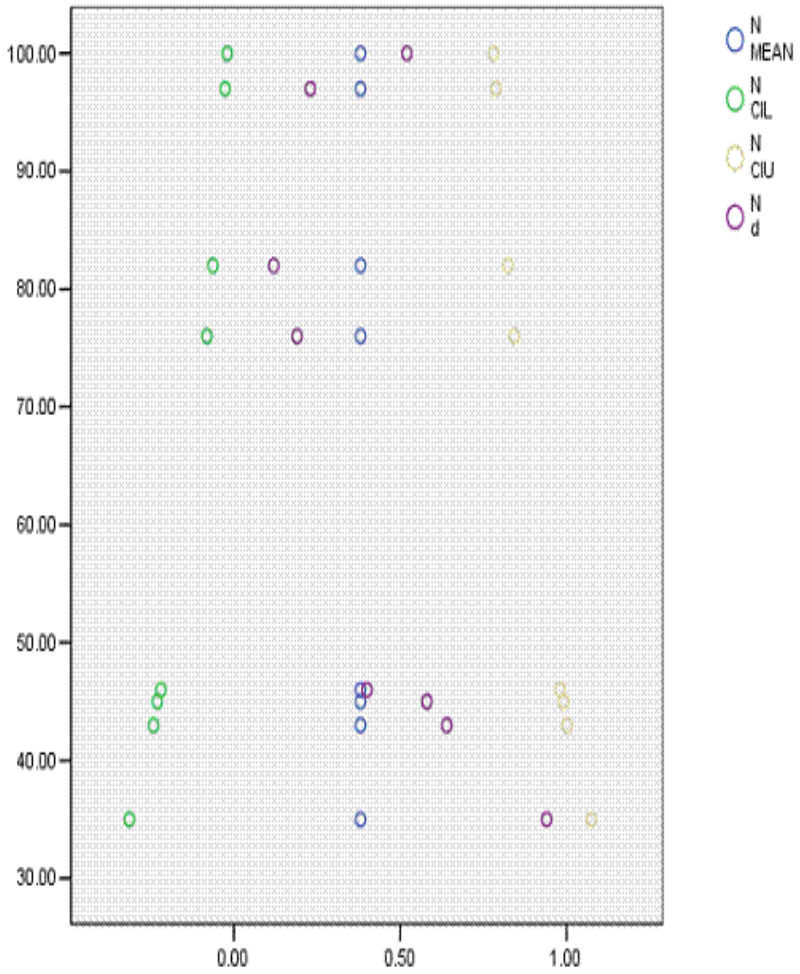
Table 8.
Rated Design Quality— L1 +Image-based Glosses versus L1 Glosses

	Author (s)	# of Control Students	# of Experimental Students	Comp Group Quality	Stat Control Quality	Design Quality	Mean Hedges Effect Size
1	Plass et al. (2003)	38	38	1	1	1	0.19
2	Wu (2002)	41	41	3	1	2	0.12
3	Akbulut (2007)	23	23	1	1	1	0.39
4	Hsu (2008)	22	21	1	4	2	0.63
5	Yoshii & Flaitz (2002)	50	50	1	4	2	0.51
6	Yoshii (2006)	47	50	1	4	2	0.22
7	Yangua (2009)	20	25	1	4	2	0.57
8	Kost et al. (1999)	18	19	1	1	1	0.92

The effects of L1 textual+image-based glosses were higher than those of L1 textual glosses alone, whereas the effect sizes of the group differences ranged from 0.12 to 0.94. However, the test of homogeneity of variance (Q-test) indicated that no statistically significant difference was detected across the effect sizes (Q=6.66, df= 7,

p=0.46). A funnel plot also reached a similar conclusion (see Figure 3). No multiple-choice glosses were used in these studies.

Figure 3.
Funnel Plot – L1+Imaged-based Glosses versus L1 Textual Glosses



Studies with the lowest level of design quality (M=0.40) were not significantly different from studies with the medium level of design quality (M=0.36) for explaining the variations among effect sizes (*QB*= 0.03, *df* =1, *p* =0. 85). Among these eight studies, only one (Kost et al., 1999) used a paper-and-pencil-based environment. No statistically significant difference (Q=3.29, *df*=1, p=0.07) was found between the group who read a

text on computer screen with glosses (M=0.34) and those who read a text from printed text materials (M=0.92). Two ranges of text-target word ratios (≤2% and between 2% and 5%) were identified. One study had text-target word ratio of ≤2% and seven studies had a ratio of between 2% and 5%. Results showed that language learners who had the text-target ratio of ≤2% (M=0.63) outperformed those who had the ratio of between 2% and 5% (M=0.36). However, the differences were not statistically significant (QB=0.94, df=1, p=0.33). No statistical significance was detected concerning the genre of reading materials (expository: M=0.48; narrative: M=0.35) among these seven studies (QB=0.42, df= 1, p= 0.52).

Discussion

Results for research question 1 (RQ) showed that the meta-analysis of the effect of L1 textual glosses on L2 incidental vocabulary learning yielded an overall positive effect size. The current meta-analyses supported results of Abraham (2008); language learners who had access to the definition of the target words obtained more vocabulary gain. As Nation (2001) suggested, the provision of the meaning of unknown words reduced the probability of inferring an inaccurate meaning of the word and fostered L2 incidental vocabulary learning.

The current meta-analyses extended the previous work by Abraham (2008) on the effect of glosses on L2 vocabulary learning by incorporating studies conducted in a paper-and-pencil-based environment with a focus on receptive vocabulary learning. Results for RQ2 from the first meta-analysis showed that studies with higher levels of design quality could more effectively detect statistically significant differences among

the groups using textual glosses on L2 incidental vocabulary learning. Therefore, to better reflect this research domain, future studies should employ groups with equivalent language proficiency levels and use a covariate of pre-treatment measure of the outcome as in repeated measures or in general linear models.

Compared to other moderator variables, such as text-target word ratios and genre of texts, the presentation medium of the text did not play a large role in L2 incidental vocabulary learning. Results for RQ 4 showed that language learners who read texts in a computer-assisted environment outperformed those who read printed texts; however, this result was not statistically significant. These results were in accordance with those of previous studies (Dillon & Gabbard, 1998; Zumbach & Mohraz, 2008) on the effect of reading comprehension.

Results for RQ 5 showed that students who read a text with a text-target word ratio of ≤2% performed best, compared to those who read text with a ratio between 2% and 5%. This result was in line with suggestions from Nation (2001) and Hu and Nation (2000). If we relate text coverage to the strands of learning from meaning-focused input and fluency development, 95% coverage is required for learning vocabulary from meaning-focused input, and language learners need to have 98-100% coverage for fluency development (Nation, 2001). Hu and Nation (2000) suggested that learners with 98% understanding of vocabulary in a text can read the text fluently by accurately guessing the meaning of unknown words. To comprehend a text and infer approximate meanings of unknown words, learners need to understand at least 95% of the words in a

text. Results from the current study supported that language learners differed in vocabulary learning with 98% and 95% coverage of a text.

Results for RQ3 from the current investigation demonstrated that language learners who were provided with multiple-choice glosses did not perform better than those who had access to single glosses. This may be because multiple-choice glosses are "inherently error prone" (Hulstijn, 1992). One of the guidelines for multiple choice question writing is that "The options [of multiple choice questions] should be plausible for the subject that does not know the correct response, permitting those that do know it to identify it and reject the others" (Moreno et al., 2006, p.70). Distractor options are usually obtained from two compatible routes: empirical and conceptual. The empirical route includes errors as distractors commonly made by learners in the given domain from the assessed data. On the other hand, the conceptual route is composed of content knowledge similar to that of the accurate answer as distractors that may be appealing to learners without sufficient knowledge. If learners are not given the correct meaning of the unknown words, they may retain the inaccurate meaning. This may interfere with language learners' L2 vocabulary development if no corrective feedback is given. As a result, advantages of multiple-choice glosses could be offset by plausible distractors, which may lead learners to choose incorrect descriptors and retain the incorrect meaning.

Koren (1999) suggested that immediate access to the correct meaning of unknown words led to correct inferences. Based on a study using the interactive hypertext program, Koren (1999) found that language learners who inferred the meaning of unknown words remembered more vocabulary in the posttest than those who had

accessed L2 textual glosses. Koren (1999) attributed these observed results of making inferences to the immediate feedback for the meaning of unknown words. After writing the answer for the meaning of target words in the reading treatment, language learners could click on the phrase "correct reply," allowing them to compare their answer with the accurate one. Mohseni-Far (2008a) also suggested that while encouraging a deeper level of lexical processing, vocabulary learning with the provision of multiple-choice glosses is discounted due to the lack of immediate feedback for learners' errors. Language learners who have access to multiple-choice glosses as well as immediate feedback can correct their guesses and enhance vocabulary development.

Results for RQ 6 showed that students who read narrative texts did not gain more vocabulary than those who read expository texts, supporting Abraham's (2008) results. Vocabulary in narrative text is usually more concrete, less content specific and less information dense than text in expository form. One possible explanation may be that the genre of reading passages did not play a large role in vocabulary learning when textual glosses were provided to language learners. The glosses provided sufficient information for language learners to read the texts and recall the meaning of newly learned words.

Results for RQ 1 of the second meta-analysis revealed that the effect of L1 textual+image-based glosses was higher than that of L1 textual on L2 incidental vocabulary learning; however, the result was not statistically significant. These results suggested that L1 textual glosses are important in L2 vocabulary learning; imaged-based glosses do not contribute much when learners also get definitions or explanations of the target words in their first language. This may be because L1 textual glosses are

straightforward and self-explanatory, which makes it easier for language learners to understand the meaning of unknown words. These results supported Yoshii's (2006) findings. Provisions of pictorial glosses may not always have great facilitative effects. As Yoshii (2006) suggested, access to pictures may not be helpful for some language learners because pictures are less precise than written definitions and are open to interpretation. Although the current investigations shed light on the effects of L1 textual glosses and L1 textual+image-based glosses on L2 incidental vocabulary learning, these results should be interpreted with caution until a sufficient body of research is available.

Conclusion

This investigation provided evidence for the facilitative effects of L1 textual glosses on L2 incidental vocabulary learning. Five moderator variables were further analyzed for variation among effect sizes: design quality of the study, multiple-choice/single glosses, computer-assisted/paper-and-pencil-based environment, text-target word ratio, and genres of reading text. As opposed to studies with a low level of design quality, studies with a high or medium level of design quality did a better job in detecting statistically significant differences among the groups.

Computer-based and paper-and-pencil-based environments did not result in different levels of vocabulary learning. Language learners who read a text with a text target-word ratio of ≤2% outperformed those who read the text with a ratio between 2% and 5%. Multiple-choice glosses and single glosses did not lead to differences in vocabulary learning. While accessing L1 textual glosses, language learners did not gain more vocabulary from reading an expository text instead of a narrative text.

Moreover, results revealed that language learners who had access to L1 textual+image-based glosses did not increase their vocabulary to a statistically significant degree over those who only had access to L1 textual glosses. Further analyses with moderator variables (design quality of the study, computer-assisted/paper-and-pencil-based environment, the text-target word ratio, multiple-choice/single glosses, and genre of reading text) also showed that no statistically significant differences existed in L2 vocabulary learning between these two groups of language learners.

Limitations and Suggestions for Further Research

Despite its comprehensive nature, this meta-analysis study has the following four limitations. First, articles without significantly different findings are more likely to be rejected by scholarly journals. If a meta-analysis only includes published articles with significant high effect sizes, a substantial skew in the distribution will occur in the research domain. Therefore, an extensive search was conducted to locate unpublished studies on the effect of L1 textual and image-based glosses on L2 incidental vocabulary learning. In spite of the effort, not all unpublished studies on the effect of L1 textual+ image-based glosses were included in the present study. Second, part of the nature of a meta-analysis is its strong reliance on the information reported in articles that are included in a meta-analysis. A lack of sufficient information, such as mean and standard deviation, to derive effect sizes will result in problems during analysis. Even after the authors of these articles were contacted, some information remained unobtainable. Thus, the present meta-analysis excluded these studies (Hulstijn, 1992; Yeh & Wang, 2003) from the original pool of the studies. Third, contrary to previous research findings, the

results of the second meta-analysis revealed that the effects of L1 textual+image-based glosses and L1 textual glosses on L2 incidental vocabulary learning did not differ significantly. A significant difference may exist between the effects of these two types of glosses on L2 incidental vocabulary learning. However, the current study could have failed to detect the significant difference because of a lack of statistical power caused by the small sample size. Therefore, more studies are needed to confirm the effects of L1 textual+ image-based glosses. Finally, the studies that were included in the second-meta-analysis did not provide the imagability of image-based glosses. Not all words are learned equally from pictures. Due to low imagability, the meanings of abstract words are less likely to be learned from images; they are also least likely to be dually coded (Sadoski & Paivio, 2001). Therefore, future studies should provide information regarding the imagability of image-based glosses to better verify the effectiveness of the visual representation of the target words.

Further research should investigate other moderate variables, such as language learners' L2 proficiency level and vocabulary size that may affect the effect of L1 textual glosses and image-based glosses on L2 incidental vocabulary learning. It might be noteworthy to consider how language learners' L2 proficiency levels and vocabulary sizes are associated with incidental vocabulary learning. However, studies included in the current meta-analyses did not provide sufficient information to make meaningful distinctions of the effects of these two variables on L2 incidental vocabulary learning. The vocabulary learning of language learners with low proficiency levels and vocabulary sizes may benefit more from L1 textual glosses than those who have higher proficiency

levels and larger vocabulary sizes. Many primary studies that were included in these meta-analyses failed to report the proficiency level and vocabulary size of the language learners under consideration. Researchers should include and report the measurements and results of learners' L2 proficiency levels and vocabulary sizes to help clarify the effectiveness of L1 textual and L1 textual+image-based glosses on L2 vocabulary learning. In addition, future studies should investigate the effect of different parts of speech, such as prefixes, suffixes, and roots, on L2 incidental vocabulary learning with the use of L1 textual+image-based glosses.

Based on the results from the current study, no differences were detected in the effect of L1 textual glosses and L1 textual+image-based glosses. Further research should investigate if the differences in vocabulary learning exist with the provision of imaged-based glosses and L2 textual+image-based glosses. The immediate effects of textual and image-based glosses were observed in L2 incidental vocabulary learning from the current investigation. However, further research is warranted to confirm if the benefits pertaining to the provisions of glosses can continue over time. These studies can help identify more effective techniques to improve retention of L2 vocabulary. It is clear that more studies on the effect of L1 textual and image-based glosses on L2 vocabulary learning are needed before a definitive conclusion can be drawn. Given these limitations, the current meta-analyses shed light on the effect of L1 textual glosses and L1 textual+image-based glosses on L2 incidental vocabulary learning.

CHAPTER IV

CONCLUSION

The purpose of this dissertation was to investigate the effect of output stimulus tasks on L2 incidental vocabulary learning. In addition, this dissertation explored the effect of L1 textual glosses and L1 textual+image-based glosses on L2 incidental vocabulary learning. Two meta-analysis studies were conducted. A total of 12 studies were included in the first meta-analysis to investigate the effect of output stimulus tasks on L2 incidental vocabulary learning. Five moderator variables were examined: design quality of the studies, types of output stimulus tasks, time on task, genre of texts, and the text-target word ratios. Results revealed that language learners who completed an output stimulus task outperformed those who merely read a text. Moreover, results were in accordance with the involvement load hypothesis, suggesting that language learners who completed a task with a higher degree of involvement load gained greater L2 vocabulary. Results also revealed that studies with a higher level of design quality or time spent on task positively correlated with L2 vocabulary learning. Learners who read an expository text gained more vocabulary than those who read a narrative text. Learners who read a text with the text-target word ratios $\leq 2\%$ did not learn significantly more vocabulary than those who read a text with a ratio between 2% and 5%.

In the second study, the effects of textual glosses and image-based glosses on L2 incidental vocabulary learning were explored. Two separate meta-analyses were conducted to investigate the effect of L1 textual glosses and the combination of L1 textual glosses and imaged-based glosses on incidental vocabulary learning. A total of

16 studies were included in the meta-analysis that compared the effect of L1 textual glosses and no glosses. Eight studies were used in the meta-analysis examining the differences between L1 textual glosses and L1 textual image-based glosses. Five moderator variables were examined in this study: design quality of the studies, computer-assisted/ paper-and-pencil-based environments, the text-target word ratios, multiple-choice/single glosses, and genres of reading texts.

Results of the meta-analysis examining the effect of L1 textual glosses and no glosses on vocabulary learning revealed that language learners who had access to L1 textual glosses gained more vocabulary than those who were not provided with any glosses. Studies with a higher level of design quality accounted for more variability in effect sizes. No significant difference in vocabulary learning was detected for learners who were provided with single or multiple-choice glosses. In addition, language learners did not show differences in vocabulary gains by using different presentation media of reading texts. Language learners who read a passage from the printed texts did not score significantly higher in a vocabulary test than those who read it on-screen. Results showed that language learners gained more vocabulary when reading a text with a text-target word ratio ≤2% as opposed to reading a text with a ratio between 2% and 5%. The genres of reading texts also did not affect learners' vocabulary learning.

Results of the meta-analysis comparing the effect of L1 textual glosses and L1 textual+image-based glosses showed that language learners gained more vocabulary when both L1 textual and image-based glosses were provided. However, these differences were not statistically significant. Further analysis of moderator variables also

reached the same conclusion. Language learners who had access to these two types of glosses did gain more vocabulary when they read a passage from a computer screen or from a printed text. However, the difference was not statistically significant. Moreover, similar results were found within each of the effects of design quality, text-target word ratios, and reading genres on vocabulary learning. Under these conditions, language learners who read a text with access to L1 textual+imaged-based glosses did not significantly gain more vocabulary than those who had access only to L1 textual glosses.

These two separate meta-analyses implied that an addition of pictures of target words did not significantly enhance participants' vocabulary learning when the L1 equivalent of the word was provided. This may be due to the fact that explanations or synonyms of the target words in the learners' first language were sufficiently explicit and simple to facilitate comprehension of unknown words. As Yoshii (2006) suggested, image-based glosses may not be beneficial for language learners when the words are provided with the definition in the language learners' first language.

Pedagogical Implications

The pedagogical implications from the findings are:

1. Teachers should incorporate tasks into the curriculum that require language learners to produce output (fill-in-the-blank vocabulary exercises, sentence completion, or composition writing) to facilitate second language vocabulary development.
2. Teachers should give language learners sufficient time to engage in output stimulus tasks. Results from this meta-analysis study suggested that language learners need approximately 40-46 minutes to engage in a task: 46 minutes to engage in a fill-in-

the-blank task, 45 minutes for a sentence-writing task, and 40 minutes for a composition writing tasks. The result of Min's (2008) study showed that 600 minutes were used for language learners to participate in a combination of different tasks. However, this result was drawn from only one study. More studies are warranted to better clarify how much time is needed to maximize the vocabulary learning via using a combination of different tasks.

3. Teachers should include glosses in reading texts to help language learners unlock the meaning of unknown words and strengthen implementation of meaning into their second language system.
4. Teachers should carefully choose a text with an appropriate text-target word ratio for the learners to maximize their vocabulary development. For instance, incidental vocabulary learning is maximized if language learners read a text with the text-target word ratio of ≤ 2% while accessing L1 textual glosses.

Limitations and Suggestions for Further Research

Scholarly journals tend to publish studies that have findings of statistically significant differences and reject those that do not (Abraham, 2008; Norris & Ortega, 2006). If a meta-analysis only includes published studies with significant differences, an inflated view with positive and high effect sizes is generated. As a result, a substantial skew may occur in the findings (Norris & Ortega, 2006). To decrease this inflated view, a thorough search was performed and attempted to include all unpublished studies on the effect of the output stimulus tasks and glosses on L2 incidental vocabulary learning in

the current investigations. Despite the effort, not all related unpublished studies were retrieved and included in the current investigation.

Additionally, the nature of meta-analysis is its high reliance on the information reported in articles. If insufficient information regarding the estimation of effect sizes, the meta-analysis researcher will encounter difficulty in getting the essential information to calculate effect sizes. Insufficient information, such as the absence of report regarding mean and standard deviations, were occurred during the process of analysis. Some of the information was still unable to obtain after contacting the authors of the original samples of studies. As a result, these studies were excluded from the current meta-analysis.

This study provided evidence favoring the use of the two most widespread techniques (output stimulus tasks and glosses) in L2 incidental vocabulary learning. However, some factors that may impacts the effect of these two techniques have not been adequately investigated. Future research should explore the role of individual differences—such as language learners' L2 proficiency levels, interest, and motivation—and how they relate to the effects of various types of output stimulus tasks and glosses in L2 incidental vocabulary learning. Additional studies on the effect of different parts of speech—nouns, verbs, adverbs, adjectives— of target words are recommended. Future research can replicate the design of this study with part of speech as a moderator variable to confirm which part of speech is better remembered with the use of output stimulus tasks.

The present study compared the effect of L1 textual glosses with L1 textual+ image-based glosses on L2 incidental vocabulary learning. No significant difference was observed in vocabulary gains for learners in the L1 textual glosses group and those in the L1 textual+image-based glosses group. A further study is warranted to investigate whether the same result emerges in the comparison of vocabulary gains for learners who had accessed the L2 textual glosses and those who were provided with L2 textual+imaged based glosses.

In addition, further studies should investigate the long term effect of these two techniques in L2 incidental vocabulary learning. The current study found that output stimulus tasks and glosses are facilitative to L2 incidental vocabulary learning, based on the immediate vocabulary posttest results. However, further meta-analysis study is warranted to examine the long term effects of output stimulus tasks and glosses to confirm whether or not the benefits of using these techniques continue or disappear over time.

REFERENCES

Abraham, L. (2008). Computer-mediated glosses in second language reading and comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning, 21(3)*, 199-226.

Akbulut, Y. (2007). Effects of multimedia annotations on incidental vocabulary learning and reading comprehension of advanced learners of English as a foreign language. *Instructional Science: An International Journal of the Learning Science*, *35(6)*, 499-517.

Allen, C., Chen, Q., Willson, V., & Hughes, J. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, *31(4),* 480-499.

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning and Technology*, *5(1),* 202-232.

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

Ben Salem, E. (2006). *The influence of electronic glosses on word retention and reading comprehension with Spanish language learners*. Unpublished doctoral dissertation, The University of Kansas, Kansas.

Bowles, M. (2004). L2 glossing: To CALL or not to CALL. *Hispania, 87,* 541-552.

Chan, S.(1999). *The Effects of word translation on college freshmen's reading comprehension and word retention*. Unpublished master's thesis, National Kaohsiung Normal University, Taiwan.

Chen, Y. (2004). The effect of marginal glosses on reading comprehension and vocabulary learning. Unpublished master's thesis, National Taiwan Normal University, Taiwan.

Chun, D., & Plass, J, (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *96,* 183-198.

Coady, J. (1993). Research on ESL/EFL vocabulary acquisition: Putting it in context. In T. Huckin, M. Haynes, & J. Coady (Eds.). *Second language reading and vocabulary learning* (pp. 16-28). Norwood, NJ: Ablex.Crailk, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Crailk, F., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory research. *Journal of Experimental Psychology: General, 104,* 268-294.

Dagut, M. (1977). Incongruencies in lexical gridding—an application of contrastive semantic analysis to language teaching. *International Review of Applied Linguistics, 15,* 221-229.

Day, R. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*,*7(2),*541-551.

Day, R., & Swan, J. (1998). Incidental vocabulary of foreign language spelling through targeted reading. *TESL Reporter, 31(1),* 1-9.

De Ridder, I.(2003).*Reading form the screen in a second language: Empirical studies on the effect of marked hyperlinks on incidental vocabulary learning, text comprehension, and the reading process.* Antwerp: Garant.

Dillon, A., & Gabbard., R. (1998). Hypermedia as an educational technology: A review of the quantitative research literature on learner comprehension, control and style. *Review of Educational Research*, 68 (3), 322-49.

Gambrell, L. (2005). Reading literature, reading text, reading the internet: The times they are a' changing. *Issues and Trends in Literac,. 58(6),* 588-591.

Gass, S. (1988). Integrating Research Areas: A Framework for Second Language Studies. *Applied Linguistics*, 9,198-217.

Hill, M. & Laufer, B.(2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *IRAL, 41,* 87-106.

Holley, F., & King, J. (1971). Vocabulary glosses in foreign language reading materials. *Language Learning*, *21 (2),* 213-219.

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

Hu, M. & Nation, P. (2000).Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13,* 403-430.

Huang, Y. (2003). *The effect of vocabulary glosses and example sentences on junior high school EFL students reading comprehension and vocabulary learning*. Unpublished master's thesis, National Cheng Kung University, Tainan, Taiwan.

Hulstijn, J. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P.J.L. Arnaud & H. Bejoing (Eds.). *Vocabulary and applied linguistics* (pp.113-125). London: Macmillan.

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51(3),* 539-558.

Hulstijn, J., 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.

Hulstijn, J., & Trompetter, P. (1998). Incidental learning of second language vocabulary in computer-assissted reading and writing tasks. In D. Albrechtsen, B. Henriksen, I. Mees., & E. Poulsen (Eds.). *Perspectives on foreign and second language pedagogy* (pp.191-202). Odense: Odense University Press.

Hsu, W. (2005). *The effect of vocabulary enhancement instruction and reading only instruction on EFL senior high school students' vocabulary acquisition in the context of a reading program*. Unpublished master's thesis, National Cheng Kung University, Tainan, Taiwan.

Hsu, M.(2008). *The effects of multimodal glosses on incidental vocabulary learning*. Unpublished master's thesis, National Taiwan Normal University, Taipei, Taiwan.

Jacobs, G. (1994). What lurks in the margin: Use of vocabulary glosses as a strategy in second language reading. *Issues in Applied Linguistics, 5 (1),* 115-37.

Jacobs, G., Dufon, P., & Hong, F.(1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, *17(1),* 19-28.

Keating, G. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, 12(3), 365-386.

Kerr, M., & Symons, S. (2006). Computerized presentation of text: Effects on children's reading of informational material. *Reading and Writing, 19,* 1-19.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58(2),* 285-325.

Knight, S. (1992).*The effect of the present or absence of dictionary access on incidental vocabulary learning for collect intermediate Spanish learners reading Spanish texts*, Unpublished doctoral dissertation, The Ohio State University, Ohio.

Ko, M. (1995). Glossing in incidental and intentional learning of foreign language vocabulary and reading. *University of Hawaii Working papers in ESL*, *13(2),* 49-94.

Koren, S. (1999). Vocabulary instruction through hypertext: Are there advantages over conventional methods of teaching? *TESL-EJ*, 4(1).Retrieve January 12th, 2010, from at http://tesl-ej.org/ej13/a2.html

Kost, C., Foss, P., & Lenzini, J.(1999). Textual and pictorial gloss: Effectiveness on incidental vocabulary growth when reading in a foreign language. *Foreign language Annals, 32(1),* 89-113.

Krashen, S. (1982). *Principles and practice in second language acquisition*. New York: Pergamon.

Krantz, G. (1991). *Learning vocabulary in a foreign language.* Göteborg, Sweden: Acta Universitatis Gothburgensis.

Kuo, F., & Chiang, H. (2006). Facilitating EFL children's vocabulary acquisition and reading comprehension with multimedia annotations. In R. Chung, H. Liou, J. Hsu, & D. Ho (Eds.), *Language and linguistics monograph series number W-7, On and off work: Festschrift in honor of professor Chin-Chuan Cheng on his 70th birthday* (pp.265-286). Taipei, Taiwan: Institute of linguistics Academic Sincia.

Labasse, B. (1999). La lisibilité rédactionelle: Fondements et perspectives. C*ommunications and Languages, 121,* 86-103.

Lan, C. (2004). *The effects of task-induced involvement load on vocabulary learning of senior high EFL students in Taiwan*. Unpublished master's thesis, National Kaohsiung Normal University, Kaohsiung, Taiwan.

Laufer, B.(1989). What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp.316-323). Clevedon: Multilingual Matters.

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook, 5,* 223-250.

Lee, K. (1995). *Bilingual texts: A study of the effect of providing L1 Chinese terms in L2 English texts on text comprehension and on English vocabulary acquisition*. Unpublished master's thesis, The University of Hong Kong, Hong Kong.

Lee, C. (2002). *The Effect of Involvement Load on English Vocabulary Acquisition in an Incidental Learning Situation*. Unpublished master's thesis, National Tsing Hua University, Hsinchu, Taiwan.

Lipsey, M. & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, Calif: Sage Publications.

Macedo-Rouet, M., Rouet, J., Epstein, I., & Fayard, P. (2003). Effects of online reading on popular science comprehension. *Science Communication*, *25 (2),* 99-128.

Min, H. (2008). EFL vocabulary acquisition and retention: reading plus vocabulary enhancement activities and narrow reading. *Language Learning, 58(1),* 73-115.

Mohseni-Far, M. (2008a). A Cognitively-oriented Encapsulation of Strategies Utilized for Lexical Development: In search of a flexible and highly interactive curriculum, *Porta Linguarum, 9,* 35-52.

Moreno, R., Martinez, R., & Muniz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2(2),* 65-72.

Nation, P. (2001). *Learning vocabulary in another language*. New York: Cambridge University Press.

Nation, P. (2006). How large vocabulary is needed for reading and listening? *The Canadian Modern Language Reviews, 63(1),* 59-82.

Norris, J.& Ortega, L.(2006). The value and practice of research synthesis for language learning and teaching. In J. Norris., & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp.3-52). Amsterdam, Netherlands: Benjamins.

Oskarsson, M. (1975). On the role of mother tongue in learning foreign language vocabulary: An empirical investigation. *ITL Review of Applied Linguistics, 27,* 19-32.

Paivio, A.(1971). *Imagery and verbal processes*. New York, NY: Holt, Reinhart & Winston.

Pitts, M., White, H., & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of clockwork orange study using second language acquirers. *Reading in a Foreign Language, 5(2),* 271-275.

Plass, J., Chun, D., Mayer, R., & Leutner., D.(2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial ability. *Computers in Human Behavior*, *19,* 221-243.

Pomplun, M., Frey S., & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, *62 (2),* 337-354.

Pretorius, E. (2006). The comprehension of logical relations in expository texts by students who study through the medium of ESL. *System*, *34*, 432-450.

Rott, S. (2004). A comparison of output interventions and un-enhanced reading conditions on vocabulary acquisition and text comprehension. *The Canadian modern language review*, *61(2),* 169-202.

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., Williams, J., & Cameron, R. (2002). The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research*, *6(3),* 183-222.

Sadoski, M., & Paivio, A. (2001). *Imagery and text: A dual coding theory of reading and writing.* Mahwah, NJ: Lawrence Erlbaum Associate.

Scherfer, P. (1993).Indirect L2-vocabulary learning. *Linguistics, 31,* 1141-1153.

Schmidt, R.(1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition, 14,*357-385.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review, 11,* 11-26.

Schmitt, N. (2008). Instructed second language learning. *Language Teaching Research*, *12 (3),* 329-363.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp.125-144). Oxford, UK: Oxford University Press.

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

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

Wa-Mbaleka, S. (2006). *A meta-analysis investigating the effects of reading on second language vocabulary learning*. Unpublished doctoral dissertation, Northern Arizona University, Arizona.

Wang, W. (2005). *The effects of vocabulary glosses on EFL vocabulary learning*. Unpublished master's thesis. Ming Chuan University, Taipei, Taiwan.

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary, *Studies in Second Language Acquisition, 19,* 287-307.

Wilhelm, J. (2000). Literacy by design: Why is all this technology so important? *Voices from the Middle, 7(3),* 4-14.

Wu, J. (2002). The effect of multimedia annotations on second language vocabulary retention. Unpublished master's thesis. Providence University, Taichung, Taiwan.

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

Yeh, Y., & Wang, C. (2003). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journal, 21 (1),* 131-144.

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology, 10(3),* 85-101.

Yoshii, M., & Flaitz, J. (2002). Second language incidental vocabulary retention: The effect of text and picture annotation types. *CALICO Journal, 20 (1),* 33-58.

Zumbach, J., & Mohraz, M.(2008). Cognitive load in hypermedia reading comprehension: Influence of text type and linearity. *Computers in Human Behavior, 24 ,* 875–887.

Supplemental Sources Consulted

Abraham, L. (2007). Second-language reading comprehension and vocabulary with media. *Hispania*, *90(1),* 98-108.

Barcroft, J. (1998, October). L2 vocabulary learning: Do sentence writing and oral repetition help? Poster session presentation at the Second Language Research Forum, Honolulu, HI.

Barcroft, J. (2000). *The effects of sentence writing as semantic elaboration on the allocation of processing resources and second language lexical acquisition*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Illinois.

Barcroft, J. (2004). Second language vocabulary acquisition: A lexical input processing approach. *Foreign Language Annals*, 37(2), 200-208.

Barcroft, J. (2006). Can writing a new word detract from learning it? More negative effects of forced output during vocabulary learning. *Second Language Research, 22 (4),* 487-497.

Bonk, W. (2000).Second Language lexical knowledge and listening comprehension. *International Journal of Listening, 14,* 14-31.

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

De la Fuente, M. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition, 24,* 81-112.

Ehri, L., & Rosenthal, J. (2007). Spellings of words: A neglected facilitator of vocabulary learning. *Journal of Literacy Research, 39(4),* 389-409.

Ellis, R. (1995). Modified input and the acquisition of word meanings. *Applied Linguistics, 16,* 409-441.

Hulstijn, J. (1993).When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variable. *The Modern Language Journal*, 77, 139-147.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal ability. *Modern Language Journal*, *78,* 285-299.

Landauer, T., & Dumais, S.(1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104,* 211-240.

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

Laufer, B., & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language*, *13*, 549-566.

Mohseni-Far, M. (2008b). In search of the best techniques for vocabulary acquisition. *Eesti Rakenduslingvistika Uhingu* Aastaraamat, 4, 121-138.

Nagy, W. (1997).On the role of context in first and second language vocabulary learning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 64-83). Cambridge: Cambridge University Press.

Paribakht, T., & Wesche, M. (1997). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*, 9-27.

Swain, M. (1984). A review of immersion education in Canada: research and evaluation studies. *ELT Documents, 119,* 35-51.

Swain, M., & Lapkin, S., (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16(3),* 371-391.

APPENDIX A

LITERATURE REVIEW

Vocabulary is considered an essential and fundamental component of second language (L2) learning (Mohseni-Far, 2008b). Learners need a large vocabulary size to achieve successful communications or perform appropriate language use in a second language. Laufer (1989) suggested that language learners should understand at least 95% text coverage, which is defined as "the percentage of running words in the text known by the readers" (Nation, 2006, p.61), for reading comprehension, whereas Hu and Nation (2000) suggested that 98% text coverage should be required to achieve good reading comprehension. Bonk (2000) found that knowing less than 90% coverage of a text resulted in inadequate reading comprehension, while more than half (60%) of those with a 95% text coverage gain a good level of text comprehension. In addition, in a review of second language vocabulary studies, Schmitt (2008) concluded that 95% text coverage is needed for adequate reading comprehension, and 98% coverage is required for language learners to reach a good level of text comprehension. Schmitt (2008) stated that 98% coverage of a text is challenging for language learners, because this text coverage presents that language learners must understand 49 words out of every 50 words to achieve this goal. Nation (2006) found that in order to reach a 98% coverage level of various written texts, such as novels or newspapers, a vocabulary size of 8,000 to 9,000 word families is required.

Although explicit teaching results in faster vocabulary gains and a higher level of vocabulary retention than learning vocabulary through reading (Schmitt, 2008), it is

difficult for foreign language learners to obtain this large vocabulary size of 8,000 to 9,000 through explicit vocabulary learning alone. Reading independently is considered the primary method to acquire new vocabulary for students above the third grade (Ehri & Rosenthal, 2007).

As a result, incidental vocabulary learning, defined as "learners acquiring new words from context without intending to do so, such as picking up new words during free reading" (Barcroft, 2004, p.201), plays a highly important role in L2 vocabulary learning.

Research (Hulstijn, 1992; Nagy, 1997; Zahar et al., 2001) supported that language learners acquire second language vocabulary from reading. However, Nagy (1997) suggested that approximately one in twenty chances occur for language learners to retain a word after a single exposure. Hill and Laufer (2003) found that learners picked up about one to five new words in a text of over 1,000 words when they are not asked to perform a word-focused activity. Based on the estimation of incidental vocabulary learning pick-up rates reported in previous research, Hill and Laufer (2003) claimed that a second language learner needs to read over eight million words of text, or approximately 420 novels, to increase vocabulary size by 2,000 words. In a review on L2 vocabulary learning studies, Schmitt (2008) concluded that incidental vocabulary has a low pick-up rate.

Laufer and Yano (2001) investigated the accuracy of language learners' self-reported understanding of word meanings in text. Language learners were instructed to read a text and answer reading comprehension questions. They were also asked to self-

assess their comprehension of target words in text. After the self-assessment, these learners were given an unexpected vocabulary test requiring them to translate the target words. Laufer and Yano (2001) compared the results of learners' vocabulary test (the objective scores) with their own self-assessment (the subjective perspectives). The comparison demonstrated that all learners over-estimated their understanding of the words, often at rates exceeding 60%. Laufer and Yano (2001) suggested that this situation may be because learners who understood the main idea of the text did not attend to the precise meaning of each single word. Another explanation was a lack of awareness of an unfamiliar word; language learners may confuse the unknown word with a word known to them. For example, learners who encountered the word "adopt" misinterpreted it as an known word "adapt," rather than an unfamiliar word whose meaning should be inferred or looked up in a dictionary. As a result, learners might not consult a dictionary for the meaning of unknown words, resulting in low L2 incidental vocabulary learning.

Laufer (2005) also suggested four reasons for a low pick-up rate in incidental vocabulary learning. First, learners who know the general concept of a text usually do not pay attention to the precise meaning of each word. Second, guessing from context is not always reliable; learners may derive an inaccurate meaning of an unknown word in text. Third, the meaning of unknown words that easily can be inferred may not yield sufficient engagement to be remembered. Fourth, repeated exposures to newly learned words are essential for language learners to integrate these words into their L2 lexical system. Language learners need to encounter a word approximately eight to ten times to

acquire it (Schmitt, 2008). According to Laufer (2005), learners need to read about one to two graded readers per week in order for words to be encountered ten times in reading; however, learners typically do not read this much. Laufer (2003) suggested "reading alone is unlikely to be the best source of vocabulary acquisition" (p.583). Schmitt (2008) determined that the following factors can foster vocabulary learning:

- Increase frequency of exposure;
- Increased noticing of the lexical items;
- Increased intention to learn the lexical items;
- A requirement to learn the lexical items (by teacher, text, syllabus);
- A need to learn/use the lexical item (for task completion or to achieve a personal goal);
- Increased amount of time spent engaging with the lexical item; and
- Amount of interaction spent on the lexical item

(Schmitt, 2008, p.339).

Schmitt (2008) suggested that greater exposure, attention, or time spent on lexical items can strengthen vocabulary learning. To enhance pick-up rates of L2 incidental vocabulary learning, many techniques—output stimulus tasks, glosses, think-aloud procedures, and increased word frequencies—have been used to draw learners' attention to the target vocabulary. Among these techniques, output stimulus tasks and glosses are widely used. The literature review aims to provide the theoretical framework for two commonly used stimulus techniques (output stimulus tasks and glosses) to facilitate second language vocabulary development. In addition, this section includes a review of research that uses these two types of stimulus techniques. This review also provides a summary of previously related meta-analyses on output stimulus tasks and glosses in second language vocabulary learning.

Output Stimulus Tasks

When language learners need to use a new word, they may acknowledge that their L2 lexical system is insufficient (De la Fuente, 2002). This may make them engage more with the target vocabulary, leading to greater vocabulary learning. Thus, output stimulus tasks, such as fill-in-the blank, sentence writing, or composition writing, are frequently used to increase L2 vocabulary development. The theoretical framework for output use includes depth of processing, output hypothesis, and involvement load hypothesis.

Depth of Processing

The degree of attention that a learner pays to a specific vocabulary word is highly associated with the extent of his or her remembering the given word. Crailk and Lockhart's (1972) depth of processing theory, "where greater depth implies a greater degree of semantic or cognitive analysis" (p.675), is widely adopted by L2 researchers. Recognition of a word may activate associations and images based on the learner's prior experience with the word. The degree to which the information is originally processed plays a dominant role in determining the likelihood that new information will be stored in long-term memory. Rott, Williams, and Cameron (2002) stated that the more attention allocated to new vocabulary, the more likely it will be learned and remembered. Learning and retention are facilitated by learners' use, reformulation, or elaboration on this new information, because these processes build connections between existing and new knowledge (Crailk & Tulving, 1975). Effective tasks should produce a "deep level of process" to contain deeper processes of new vocabulary.

Output Hypothesis

Swain (1984) describes a scenario in Canada, where students in a French-language immersion program were provided a considerable amount of comprehensible input. These students achieved fluency in French and were generally able to communicate with their bilingual teacher and peers. However, when their mastery reached the point at which they could understand and be understood, their second language development rate seemed to slow down. These immersion students' expressive performance was much less satisfactory than that of native French speakers of same age. They showed lower precision in their vocabulary use, less accuracy in pronunciation, and less mastery of grammar. These results from the immersion program provide strong counter evidence for a focus on comprehensible input. Therefore, Swain contends that mere understanding of new language forms is not sufficient in second language learning. Learners also need to be provided opportunities to produce the new language.

Similarly, Gass (1988) suggests a metaphorical characterization of ways new lexical elements may be integrated into learners' mental lexicons over time and made available for communication. The process consists of five stages—apperceived input, comprehended input, intake, integration, and output. In the first stage, language learners are exposed to a considerable amount of second-language data (i.e., input), but do not use all of the input. Some of the data will pass by language learners, and some may not. The first stage, apperceived input, is defined as language learners' understanding of the initial data. Apperception is referred to as the process of understanding by which "newly observed qualities of an object are related to past experiences" (Gass, 1988, p.201);

namely, prior knowledge is associated with the selection of noticed material. High or low frequency of occurrence, salience, and the availability of prior knowledge may affect what is noticed from the available sensory information. The second stage is comprehended input. Unlike Krashen's use of the term, Gass (1988) argues that this is a multi-staged concept that has different levels of comprehension, such as general meaning, semantics, and detailed structural analyses. For instance, language learners can understand the component parts of an utterance and grasp their syntactic or phonological pattern. Not all comprehended input will move to the next stage of intake, defined as a process of attempted integration of linguistic information, mediating between target language input and the learners' existing internalized rules. Intake is not solely a subset of input, but a distinct phenomenon. The following stage is integration, which is an outcome of the intake process. Integration involves changes to the internalized second-language rule system based on the new information. Output, the last stage, plays a pivotal role in transforming comprehended input into intake, and learners' output can represent their integrated knowledge. In addition, output forces learners to reflect on adequate and specific target-language knowledge for their intended messages. Output enhances language development. Therefore, teachers should provide language learners with opportunities for producing output and meaningful use of the target language.

Production of the target language enhances fluency. Output of the target language helps learners improve fluency, but does not guarantee increased accuracy (Schmidt, 1992). Swain (1995) proposes three output functions that improve accuracy more than fluency. She contends that output serves three functions in language learning:

noticing, hypothesis testing, and a metalinguistic function. First, output production helps language learners notice gaps in their language proficiency while they attempt to produce the target language. Output helps learners notice what they intended to say but could not say precisely or comprehensively. When learners notice the gap and realize they do not know the word in the L2, they return to the text with more focused attention and fill the newly identified gap. As Swain and Lapkin (1995) contend, noticing may take place in response to either internal or external feedback, which helps facilitate the generation and selection of alternatives through a deliberate inspection of complex thinking. If learners cannot come up with a solution, they may return to the text with even greater attention to search for relevant input. Second, even when external feedback is unavailable, learners can use output to test out new language forms (hypotheses). Some errors in speech or writing underscore the hypothesis about how the target language works and its comprehensibility. To test the hypothesis, learners have to produce output in verbal or written form. The production of either form can help language learners test their own hypotheses. The output itself becomes a hypothesis that reflects the learner's best guess about how the language should be used. In general, teachers do not ask learners to describe their hypotheses; instead, teachers infer them from the output provided. However, under certain conditions, learners will both reveal their hypotheses and reflect on them through their use of language. This "level" of output represents the metalinguistics functions by which language use reflects the language itself and allows learners to control and internalize it (Swain, 1995).

The Involvement Load Hypothesis

Compared to input, output requires more mental effort on the part of learners. Learners can pretend to comprehend while reading, but they cannot do so while speaking or writing. The process of vocabulary growth may reach a plateau unless learners are given the opportunity to put this vocabulary to use and develop skills in its use (Nation, 2001). The involvement load hypothesis proposed by Laufer and Hulstijn (2001) explains the extent of cognitive processing required of a language learner to perform a given task. Laufer and Hulstijn (2001) mention that the degree of the mental effort or involvement load required by a given task determines word learning and retention.

Task-induced involvement contains three components: need, search, and evaluation. Need is related to the motivational, noncognitive dimension of the task involvement. Need is defined as whether or not knowledge of a new word is required to complete a task. Need is considered moderate when it is extrinsically imposed by other agents, such as teachers or task completion, such as when a learner needs to complete reading comprehension questions that require knowledge of the unknown word. Need is strong when it is intrinsically imposed by the learner, such as when the learner feels the need to find out the meaning of the unknown word in order to write a better composition.

Search and evaluation are associated with the cognitive dimension of the task involvement. Both search and evaluation involve allocating attention to a word form and its meaning. Search is present when the learner attempts to identify the meaning of an unknown word using other sources, such as a dictionary or teachers and peers. Search is absent when the searching attempt is not made. Evaluation is defined as the decision-making process during tasks, such as “a comparison of a given word with other words, a

specific meaning of a word with its other meanings, or comparing the word with other words in order to assess whether a word … does or does not fit its context" (Laufer & Hulstijn, 2001, p.14). Evaluation can be moderate or strong. Evaluation is considered moderate when the learner is required to evaluate a particular word or words to fit a given context. On the other hand, evaluation is considered strong when the words being evaluated must combine the newly learned word with additional words in an original text.

Studies (Barcroft, 2004; Barcroft, 2006; Keating, 2008; Laufer, 2003; Tu, 2004) have shown that input-plus-output tasks benefit L2 vocabulary development more than input-task only. Hulstijn and Trompetter (1998) found that language learners in the composition writing group gained more vocabulary than those who were in the reading-only group. Moreover, Paribakht and Wesche (1997) found an additional word-focused exercise is more effective than a reading task only on vocabulary learning.

Laufer (2003) compared the effects of three types of output stimulus tasks (fill-in-the-blank, sentence writing, and composition) with the effect of reading only on L2 vocabulary learning. Laufer (2003) found that compared to learners in the reading-only group, learners in an output stimulus task were more likely to remember words, and suggested that if learners overestimated their understanding of the words, the output tasks help them correct this misperception. Output stimulus tasks force learners to notice unfamiliar words, making them aware of their insufficient knowledge. As a result, language learners performing output stimulus tasks have a more accurate estimate of their knowledge of the target vocabulary as opposed to the reading-only group. In

addition, output stimulus tasks require a greater extent of involvement load, because learners are asked to *use* newly learned words accurately in output stimulus tasks compared to the reading-only group (Laufer & Hulstijn, 2001). Inferring the meaning of unknown words in text may not require the same extent of effort as output use.

Similarly, Tu (2004) investigated the effect of task involvement on L2 vocabulary development. In Tu's (2004) study, participants were Taiwanese high school students randomly assigned to either the control or experimental groups. In addition to reading a text, the experimental groups were asked to complete one of the tasks: reading comprehension plus fill-in-the-blank or writing a composition with the target words. The control group was required to read the same text as the experimental groups. After the experiment, all control and experimental groups were asked to read the text and complete a vocabulary posttest that required them to provide Chinese equivalents or English explanations for the target words. The results revealed that participants who were in the writing composition group earned the highest score in the immediate vocabulary posttest, followed by reading comprehension plus fill-in-the-blank, and the control group.

In Keating's (2008) study, 79 undergraduate students were randomly assigned to one of the following groups: reading, reading comprehension plus fill-in-the-blank, and sentence writing with the target words. In addition to reading a text, each group completed the assigned task. Immediately after the experiment, each group was instructed to complete a passive vocabulary recall test. Participants in the sentence

writing task group outperformed the other two. Participants in reading comprehension plus fill-in-the-blank earned higher scores than those in the reading-only group.

In contrast, Barcroft (1998) stated that language learners writing new words in sentences exhausted processing resources; a strong detrimental effect on L2 vocabulary learning existed for learners in the sentence writing group compared to the group that was exposed only to the target words. All participants were exposed to 24 Spanish word-picture pairs that were displayed individually on a television screen. The word-picture pairs were presented to learners in the no sentence writing group without additional tasks. In the sentence writing group, the word-picture pairs were presented to learners on a screen while they wrote sentences for each target Spanish word. All participants were given an unexpected immediate vocabulary posttest. Results of the vocabulary test revealed that language learners in the sentence writing group did not outperform those who in the no sentence writing group. Barcroft (2000) suggested that sentence writing may be less effective than only exposure to the words. While learning new words, learners need to allocate processing resources both to encoding new L2 word forms and establishing the form-meaning connections. If language learners exhaust their processing resources while performing output tasks, their ability to perform these two processes (encoding new L2 words and developing the form-meaning connections) might be hampered. Barcroft (2000) suggested that when a task requires more than the availability of processing resources, semantic processing may facilitate semantic learning, rather than word form learning. This might be due to the uneven distribution of the semantic and form components of learning.

Previous Meta-analysis on Output Stimulus Tasks

Wa-Mbaleka (2006) conducted a meta-analysis on the effectiveness of depth of processing while reading on L2 incidental vocabulary learning. In Wa-Mbaleka's (2006) study, eight levels of depth of processing were included: "Level 1: reading authentic texts with no dictionaries/glosses; Level 2: reading modified texts or from class textbooks OR reading plus other L2 activities [e.g., discussion or think-aloud, or recall activities for target vocabulary]; Level 3: reading with target words highlighted; Level 4: reading with glosses or dictionaries; Level 5: reading +some productivities ways of using news words; Level 6: writing; Level 7: reading+ direct learning of L2 words; and Level 8: direct instruction/learning of L2 words" (p.207). However, the results did not reveal significant differences between these levels. Wa-Mbaleka (2006) concluded that the considerable heterogeneity in the effect sizes, as a result of too many levels of processing (eight), did not lead to meaningful results. Therefore, Wa-Mbaleka (2006) suggested that future research use simpler classifications to yield more meaningful results. To summarize, some of the above-mentioned research (Barcroft, 2004; Barcroft, 2006; Keating, 2008; Laufer, 2003; Tu, 2004) provided evidence in favor of depth of processing, output hypothesis, and involvement load hypothesis. Language learners benefit from actively engaging in tasks that require language production or involvement. Other research (Barcroft, 2000) suggested that producing output may deplete processing resources and lead to less vocabulary gain. A meta-analytic view of this issue is warranted to advance our understanding of the overall effects of different output stimulus tasks on L2 incidental vocabulary learning. Further analyses on other

moderators, such as genres of texts, text-target words ratios, and treatment length, are needed to help vocabulary educators and English teachers develop a better method to improve incidental vocabulary learning with the use of output stimulus tasks.

Glosses

In addition to output stimulus tasks, glosses are another common technique to enhance vocabulary learning. While reading independently, language learners often apply guessing strategies to unlock the meaning of unknown words. Guessing from context is one of the most frequent and preferred strategies for language learners while encountering unknown words in reading (Nation, 2001). Although language learners tend to better retain the inferred meanings of newly learned words than words with vocabulary glosses (Hulstijn, 1992; Laufer & Hulstijn, 2001), they do not always derive accurate meanings if the contextual information is insufficient. To decrease the possibility of incorrect inferences, glosses—a word provided with its definition or synonym in the language learner's first or second language—is widely used. According to Nation (2001), the advantages of glosses are threefold. First, glosses enable language learners to comprehend authentic texts; otherwise, a text may be difficult to understand at the present proficiency level. Second, language learners can reduce the possibility of inferring inaccurate meanings of unknown words while accessing glosses. Third, glosses help language learners notice unfamiliar words. Noticing unfamiliar words in text is essential in vocabulary learning. "Learning is noticing" (Nation, 2001, p.63) means paying attention to a vocabulary item to accomplish a great level of mastery in second language; students must notice a word and become aware of it as a useful language item.

Noticing may enhance the processes of language acquisition. Ellis (1995) agrees that noticing plays a fundamental role in second language acquisition. He suggests, "No noticing, no acquisition" (p.89). In addition, learners gain more automaticity in their vocabulary development because through glosses, they become less reliant on teachers for explanations.

Glosses are widely used to help language learners advance their vocabulary development. The provision of glosses can strengthen vocabulary learning from texts. Different forms of glosses—textual, image-based, audio, or a combination—are used to enhance L2 vocabulary learning. Among these types of glosses, textual and imaged-based glosses have been commonly used to facilitate L2 vocabulary learning. Textual glosses provide a straightforward explanation or definition of the unknown word in either the learner's first or second language, whereas image-based glosses provide language learners with a visual representation regarding the meaning of the unknown word. Glosses of text or imagery are used to integrate the newly learned words in language learners' L2 lexical system.

Many studies (Bowles, 2004; Chen, 2004; Hulstijn, 1992, 1993; Jacobs et al., 1994; Knight, 1994; Watanabe, 1997; Yoshii, 2006) have investigated the effectiveness of L1 and/or L2 textual glosses in L2 incidental vocabulary learning. In Jacobs et al.'s (1994) study, participants enrolled in a Spanish language program at a university were randomly assigned to either the control, English (L1) glosses, or Spanish (L2) glosses groups. All participants were instructed to read an authentic Spanish article. In addition, some groups received English glosses or Spanish glosses while reading. After reading,

participants completed vocabulary recall and translation tests as well as a questionnaire investigating their perception of vocabulary glosses: namely, their preference of vocabulary glosses or no glosses, preference regarding the language of glosses, and the location of glosses in the text. Jacobs et al. (1994) found no significant difference between the effectiveness of L1 and L2 textual glosses on the recall test. On the other hand, participants with access to glosses performed better on the translation test than those lacking access. Jacobs et al. (1994) also found that almost all (98.7%) participants preferred to have vocabulary glosses; the most favored location for glosses was in the margin of the text. Almost half (47%) responded that they preferred English glosses, and over half (52%) preferred "Spanish glosses if they could understand them but English glosses if they could not" (Jacobs et al., 1994, p.26).

Researchers (Chen, 2004; Yoshii, 2006) found no significant difference between the effectiveness of L1 and L2 textual glosses. These results suggested that vocabulary glosses in language learners' first or second language did not differ greatly in enhancing their vocabulary learning. However, other researchers (Krantz, 1991; Oskarsson, 1975; Scherfer, 1993) found that language learners who had accessed L1 textual glosses outperformed those who were provided L2 textual glosses. Thus, for some language learners, vocabulary glosses in their first language can more effectively enhance L2 vocabulary development. Schmitt (2008) suggested that the provision of L1 translations for target words helps learners develop the initial link between the word meaning and its form. An L1 translation provides an easy access to establish the initial connection between the word form and its meaning, because while processing the new words,

language learners will directly connect new L2 words with the L1 words that have a corresponding, pre-existing L1 concept in memory (Dagut, 1977; Schmitt, 2008). After the connection is established, more cognitive resources will be freed up. As a result, language learners can attend to more L2 word forms and contextualized word knowledge.

Multiple-choice or single textual glosses

A word with a derived meaning is more likely to be retained in an L2 lexical system than a word with a glossed meaning (Nation, 2001). However, learners may infer incorrect meanings of an unknown word if the text provides insufficient information. On the other hand, while a single textual gloss provides language learners with the accurate meaning of an unknown word, it fails to offer opportunities for learners to infer the meaning of the word on their own. Thus, Hulstijn (1992) introduced multiple-choice glosses that allow learners to choose the correct meaning of a word from multiple options. In addition, while reading a text with multiple-choice glosses, learners are less likely to establish an incorrect form-meaning connection compared to when reading an unglossed text. As opposed to reading a text with single textual glosses or a text without any glosses, access to multiple-choice glosses in text engages learners at a deeper level of word processing. First, language learners have to conduct a search for meaning by referring to the gloss options. Second, they need to evaluate the different meanings (options in the multiple-choice glosses) and decide the appropriate meaning (the option) that best fits the context of the unknown word. Thus, the provision of multiple-choice

glosses may establish a stronger connection between word form and meaning than single textual glosses or an unglossed text.

Studies have had mixed findings with regard to the effect of multiple-choice and single textual glosses on L2 vocabulary learning. Hulstijn (1992) and Rott (2005) found that language learners with access to multiple-choice glosses in text developed robust form-meaning connections. Hulstijn (1992) investigated the impact of providing language learners with single textual glosses or multiple-choice glosses on L2 incidental vocabulary learning. Participants were assigned to three groups: reading with single textual glosses, with multiple-choice glosses, and without any glosses. Hulstijn (1992) found that, compared to those who had accessed the meaning of single textual glosses, language learners in the multiple-choice glosses group tended to remember more the form and meaning of an unfamiliar word (high mental effort) in text. Moreover, language learners were more likely to derive an incorrect meaning of an unknown word in a text while reading a text without glosses as opposed to the other two groups with access to either single or multiple-choice glosses. Rott (2005) also compared learners' L2 vocabulary gains from using multiple-choice and single textual glosses. A total of ten native speakers of English with German as a foreign language participated in Rott's (2005) study. A tape-recorded think-aloud procedure was used to investigate language learners' cognitive mechanisms that established and strengthened the link between word form and meaning (Rott, 2005). These learners received a reading text with either multiple-choice or single textual glosses. They were also asked to read the text for comprehension and verbalize their thinking while reading. Rott (2005) suggested that the

interpretations of learners' word processing strategies provided insight into their metal activities. Rott (2005) found multiple-choice glosses yielded stronger form-meaning connections as opposed to single textual glosses based on the think-aloud data.

However, other researchers (Wang, 2005, Watanabe, 1997) found no significant difference between language learners who used multiple-choice and single textual glosses. Wang (2005) found that language learners who had accessed single textual glosses performed better than those who were provided multiple-choice glosses in the first posttest seven days after treatment, but not in the delayed posttest given two weeks after treatment. Wang (2005) suggested that the provision of single textual glosses in L2 incidental vocabulary learning seems to be more effective than multiple-choice glosses in the short term, but this effect decreases over time. Learners who were provided with single textual or multiple-choice glosses outperformed those who were not provided any glosses in a vocabulary posttest. Similarly, Watanabe (1997) found that the groups that received single glosses or multiple-choice glosses outperformed the group without access to any glosses. However, no statistically significant difference was detected between groups using single textual and multiple-choice glosses.

Dual Coding Theory

Image-based glosses are also frequently used to enhance second language learners' vocabulary learning. As words are learned, they become integrated into a network of semantic connections to other words (Ehri, & Rosenthal, 2007; Landauer & Dumais, 1997). Images as well as verbal definitions and interconnections contribute to the representations of words in memory (Sadoski, 2005). Dual coding theory (DCT)

proposes a concept of how information is stored in one's brain. In DCT, coding is referred to as "the ways the external word is captured in those internal forms" (Sadoski & Paivio, 2001). Information is stored in two different but complementary systems—verbal and imagery systems. The verbal system processes information associated with language, namely, a word's pronunciation and spelling. On the other hand, the imagery system deals with nonverbal information, including production of scene analysis, generation of mental images, and representation of sensory modalities. Dual coding theory suggests that language learning is best achieved when language materials apply both verbal and imagery subsystems in the learning process.

In addition, technological advances have significantly changed the methods of language learning and teaching. Advanced computer techniques have diversified the use of glosses in L2 vocabulary learning. Language learners use a wider array of glosses—texts, pictures, animations, videos, or sounds—with the integration of computer-based or multimedia programs in vocabulary development.

Research has adopted a variety of gloss forms in L2 vocabulary learning (Abraham 2007; Akbulut, 2007; Al-Seghayer 2001; Chun & Plass, 1996; Yoshii & Flaitz, 2002). Results have shown that learners who are provided glosses outperform those who are not. Therefore, interest in glosses has been shifting from the effectiveness of glosses to the types of glosses most effective in L2 incidental vocabulary learning. Advanced technology provides greater availability of computer use, which facilitates image integration in language teaching material. Thus, many researchers have investigated the effectiveness of image-based and/or textual glosses (Abraham, 2007;

Akbulut, 2007; Al-Seghayer 2001; Chun & Plass, 1996; Yoshii & Flaitz, 2002, Yoshii, 2006).

Chun and Plass (1996) investigated the effectiveness of multiple types of glosses—text, picture, and video—on vocabulary learning. Participants were German-language learners enrolled at Stanford University and the University of California, Los Angeles. Participants were instructed to read a short story for the purpose of gauging reading comprehension. While reading, all 15 target words were provided with glosses; five had only textual glosses, five had picture and textual glosses, and five had video and textual glosses. After two weeks, the participants were asked to complete an unexpected vocabulary test. The results revealed a significant difference in the effectiveness of these glosses types. Participants learned the most words with picture-plus-textual glosses, followed by words that were provided with video-plus-textual glosses, and then only with textual glosses.

Yoshii and Flaitz (2002) explored the impact of different multimedia glosses—text, picture, and video—on L2 vocabulary learning. In Yoshii and Flaitz's (2002) study, participants were ESL students enrolled in English language institutes at various universities in Florida. Participants were informed that the purpose of the study was reading comprehension and then were randomly assigned to three groups: text-only, picture-only, and text-and-picture. All participants were instructed to read the text for comprehension. While reading, the text-only group received textual glosses for the target words, the picture-only group was provided pictorial glosses, and the text-and-picture group had a combination of textual and pictorial explanations for the target words. After

completing the reading, all three groups were given a reading comprehension task with three unexpected vocabulary tests: definition supply, picture recognition, and word recognition. Results showed that the text-and-picture group outperformed the other two groups on all three tests.

Likewise, Yoshii (2006) investigated the effectiveness of various types of glosses on incidental vocabulary learning. In Yoshii's (2006) study, participants were randomly assigned into four groups with different glosses types: L1-text-only, L2-text-only, L1-text-plus-picture, and L2-text-plus-picture. While reading the text, each group was given different types of glosses. After reading the text, the students completed two unexpected vocabulary immediate posttests: a definition supply test and a recognition test. In the definition-supply vocabulary test, Yoshii (2006) found a significant difference between the L1 and L2 textual gloss groups and between the picture and no-picture groups. Among these four groups, the L1-plus-picture group performed the best. However, no significant difference was detected on the recognition vocabulary test between the L1 and L2 textual groups and between the picture and no-picture groups.

In a recent study, Akbulut (2007) also explored the effectiveness of types of multimedia glosses on L2 incidental vocabulary acquisition. Participants were randomly divided into three conditions: textual definition, picture, and video-and-definition. While reading the text, participants were provided with the different types of glosses, after which they took three unexpected vocabulary tests—form recognition, meaning recognition, and meaning production. In the form recognition test, participants were provided with a checklist and asked to indicate the words they remembered from the

text. In the meaning recognition test, students were asked to choose the correct meaning in the form of multiple-choice questions for the target words. In the meaning production test, students were instructed to write the L1 equivalents or definitions of target words. Results revealed that the combination of visual and textual glosses led to better learning on all three vocabulary tests. In addition, on all three tests, both the picture and the video-and-definition groups outperformed the definition-only group.

Abraham (2007) compared the effectiveness of different kinds of multimedia glossed words in L2 vocabulary learning. Participants were randomly assigned to one of three groups: control, choice lookup, and forced lookup. In the control groups, participants were not provided any types of glosses. Learners in the choice lookup group had access to L1+ L2 textual and pictorial glosses, whereas learners in the forced lookup group were required to consult all L1+L2 textual and picture glosses. Results showed that learners in both choice lookup and forced lookup groups outperformed the control group on the vocabulary posttest. However, no significant differences were detected on the vocabulary posttest between the choice and forced-lookup groups.

In summary, the above-mentioned studies suggest that findings regarding the effects of the combination of textual definition and visual presentation are far from conclusive. The use of meta-analysis has been proposed as a solution to clarify the research picture. This format can account for a causal relationship between variables by statistically combining studies' findings and analyzing them from a broader perspective than an individual study.

A meta-analysis conducted by Abraham (2008) explored the effect of computer-mediated glosses in L2 vocabulary learning while reading, using a total of six studies. An effect size of 1.4 for the immediate vocabulary posttest and 1.25 for the delayed vocabulary posttest was found between groups with and without the provision of vocabulary glosses. In addition, the results indicated that glosses were most effective for students at the intermediate proficiency level but least effective for those at the beginner proficiency level; however, the difference was not significant. With regard to the different genres of text, the effect size of 1.52 for expository and 1.28 for narrative texts was detected on the immediate vocabulary posttest, but the effect was not significantly different. As Abraham (2008) suggested, these results were tentative and should be interpreted with caution because of the small number of studies included in the meta-analysis.

Although this meta-analysis (Abraham, 2008) provides a broad understanding of the effect of textual glosses on L2 vocabulary learning, some issues remain. Abraham's (2008) research did not include studies conducted in paper-and-pencil settings, but rather focused only on studies in computer-mediated environments. The magnitude of effect on textual glosses effect may vary in these two types of settings. Furthermore, many moderators, such as the effect of the ratio of text-target word and multiple-choice glosses, were not included in the previous study. In addition, much research (Al-Seghayer, 2001; Chun & Plass, 1996; Kost, Foss, & Lenzini, 1999; Yoshii & Flaitz, 2002) has generated mixed results on the effects of imaged-based glosses on L2 incidental vocabulary learning. Further research is, therefore, warranted to address the

effects of these two types of glosses on L2 incidental vocabulary learning. Extending the previous meta-analysis studies (Abraham, 2008; Wa-Mbaleka, 2008), the current investigation aims to provide broader and richer meta-analytic views of the effects of output tasks and glosses on L2 incidental learning with an emphasis on vocabulary.

APPENDIX B

CODING SHEET FOR THE FIRST META-ANALYSIS STUDY [VARIABLE NAMES IN BRACKETS]

__ __ __ __ 1. Study ID number [STUDYID]

__ 2. Type of publication [PUBType]
1. journal article
2. thesis or dissertation
3. book/book chapter
4. technical report
5. conference proceeding
6. other (specify):_____

__ __ 3. Publication year [PUBYear]

Sample Descriptors

___ 4.Mean age [MAge]

___ 5. First language of the learners [L1]
1. Korean
2. Japanese
3. English
4. Dutch
5. Hebrew
6. Khmer
7. other (specify):_____

___ 6. L2 of the learners [L2]
1. English
2. Spanish
3. German
4. French

___ 7. L2 proficiency level [L2Pro]
1. beginner
2. intermediate
3. upper intermediate
4. advanced
5. cannot tell

___ 8. School of Students [Sch]
1. middle school
2. high school
3. university
4. cannot tell

___ 9. Years of studying English[YStuL2]
1. >10 years
2. > 11 & <15 years
3. >16 years

4. cannot tell

Research Design Descriptors

___ 10. Type of assignment to condition [AssCon]
1. random
2. nonrandom
3. cannot tell

___ 11. Was the equivalence of the groups tested at pretest? [PreEqu]
1. yes
2. no
3. cannot tell

___ 12. Pretest is conducted [Pretest]
1. yes
2. no

___ 13. Interval of pretest and treatment [IntraPT]
1. immediate
2. 3 days
3. 5 days
4. 1 week
5. 2 weeks
6. cannot tell

___ 14. Interval of treatment and 1st delay posttest [IntraTP]
1. immediate
2. 1 day
3. 3 days
4. 14 days
5. 4 weeks
6. other

___ 15. Interval between 1st and 2nd delay posttests [IntraPP]
1. 1 week
2. 2 weeks
3. 4 weeks
4. 5 weeks
5. other

___ 16. Treatment group sample size [Ne]

___ 17. Control group sample size [Nc]

___ 18. Total sample size [N]

Nature of the Treatment Descriptors

___ 19. Number of target words [Ntw]

___ 20. Text lengths [Textlegth]

___ 21. Ratio of the target word and the text length [Rtwtex]

___ 22. Range of the ratio of the target words and the text length [RangRatio]
1. less than 2%
2. greater than 2% but less than 5 %

3. greater than 5%

___ 23. Number of distractors in the pretest [Ndistra]

___ 24. Treatment duration in minutes [TreatDur]

___ 25. Selection of the text [SeleText]

1. interesting
2. topic familiarity
3. difficulty level

___ 26. Types of tasks [TofTask]

1. fill-in-the-blank
2. sentence writing
3. composition
4. other_____

___ 27. Computer-assisted [ComAss]

1.yes
2.no

___ 28. # of exposures to the target words [EXPOSU]

___ 29. Part of Speech of TW [SpeechTw]

1. verb
2. noun
3. adverb
4. mixed
5. cannot tell

___ 30. Frequency of TW [FreqTW]

1. low frequency
2. High frequency
3. academic words
4. cannot tell

___ 31. Pilot of the target words [PilotTW]

1. yes
2. no

___ 32. Genres of the reading text [TText]

1. expository
2. narrative
3. cannot tell

___ 33. Authenticity of the text [AuthenText]

1. authentic
2. authentic and revised
3. author wrote
4. cannot tell

___ 34. Flesch-Kincaid grade level Readability [Readability]

Effect Size Data

___ 35. Type of effect size [TypeES]

1. means and standard deviations

2. *t-test* or *f-value*
3. chi-square (*df*=1)
4. other_____

__ __ __ __ 36. Page number where effect size data found [PEs]

Means and Standard Deviations

__ __ __ __ 37a. Mean of Treatment group [MT]
__ __ __ __ 37b. Mean of Control group [MC]
__ __ __ __ 37c. Standard Deviation of Treatment group [SDT]
__ __ __ __ 37d. Standard Deviation of Control group [SDC]

APPENDIX C

CODING SHEET FOR THE SECOND META-ANALYSIS STUDY [VARIABLE NAMES IN BRACKETS]

__ __ __ __ 1. Study ID number [STUDYID]

____ 2. Type of publication [PUBType]

1. journal article
2. thesis or dissertation
3. book/book chapter
4. technical report
5. conference proceeding
6. other (specify):_____

__ __ 3. Publication year [PUBYear]

Sample Descriptors

___ 4.Mean age [MAge]

___ 5. First language of the learners [L1]

1. Korean
2. Japanese
3. English
4. Dutch
5. Hebrew
6. Khmer
7. other (specify):_____

___ 6. L2 of the learners [L2]

1. English
2. Spanish
3. German
4. French

___ 7. L2 proficiency level [L2Pro]

1. beginner
2. intermediate
3. upper intermediate
4. advanced
5. cannot tell

___ 8. School of Students [Sch]

1. middle school
2. high school
3. university
4. cannot tell

___ 9. Years of studying English[YStuL2]

1. >10 years
2. 11 & <15 years
3. >16 years

4. cannot tell

Research Design Descriptors

___ 10. Type of assignment to condition [AssCon]
1. random
2. nonrandom
3. cannot tell

___ 11. Pretest was conducted [Pretest]
1.yes
2.no

___ 12. Was the equivalence of the groups tested at pretest? [PreEqu]
1. yes
2. no
3. cannot tell

___ 13. Interval of pretest and treatment [IntraPT]
1. immediate
2. 3 days
3. 5 days
4. 1 week
5. 2 weeks
6. other

___ 14. Interval of treatment and 1st delay posttest [IntraTP]
1. immediate
2. 1 day
3. 3 days
4. 14 days
5. 4 weeks
6. other

___ 15. Interval between 1st and 2nd delay posttests [IntraPP]
1. 1 week
2. 2 weeks
3. 4 weeks
4. 5 weeks
5. other

___ 16. Treatment group sample size [Ne]
___ 17. Control group sample size [Nc]
___ 18. Total sample size [N]

Nature of the Treatment Descriptors

___ 19. Number of target words [Ntw]
___ 20. Text lengths [Textlegth]
___ 21. Ratio of the target word and the text length [Rtwtex]
___ 22. Range of the ratio of the target words and the text length [RangRatio]

1. less than 2%
2. greater than 2% but less than 5 %
3. greater than 5%

___ 23. Number of distractors in the pretest [Ndistra]

___ 24. Treatment duration in minutes [TreatDur]

___ 25. Selection of the text [SeleText]

1. interesting
2. topic familiarity
3. difficulty level

___ 26. Computer-assisted [ComAss]

1. yes
2. no

___ 27. # of exposures to the target words [EXPOSU]

___ 28. Part of Speech of TW [SpeechTw]

1. verb
2. noun
3. adverb
4. mixed
5. cannot tell

___ 29. Frequency of TW [FreqTW]

1. low frequency
2. high frequency
3. academic words
4. cannot tell

___ 30. Pilot of the target words [PilotTW]

1. yes
2. no

___ 31. Genre of the reading text [TText]

1. expository
2. narrative
3. cannot tell

___ 32. Authenticity of the text [AuthenText]

1. authentic
2. authentic and revised
3. author wrote
4. cannot tell

___ 33. Flesch-Kincaid grade level Readability [Readability]

___ 34. Single/multiple-choice glosses [SingMul]

1.single glosses
2.multiple-choice glosses

Effect Size Data

___ 35. Type of effect size [TypeES]

1. means and standard deviations

2. *t-test* or *f-value*
3. chi-square (*df*=1)
4. other_____

__ __ __ __ 36. Page number where effect size data found [PES]

Means and Standard Deviations

__ 37a. Mean of Treatment group [MT]
__ __ __ __ 37b. Mean of Control group [MC]
__ __ __ __ 37c. Standard Deviation of Treatment group [SDT]
__ __ __ __ 37d. Standard Deviation of Control group [SDC]

VITA

Name: Shu-fen Huang

Address: Department of Curriculum and Instruction
c/o Dr. Zohreh R. Eslami
Texas A&M University
College Station, TX 77843-4232

Email Address: shufen.hhuang@gmail.com

Education: B.A., Finance, Chaoyang University of Technology, 2001
M.Ed., Curriculum and Instruction, Texas A&M University, 2005
Ph.D., Curriculum and Instruction, Texas A&M University, 2010