THE UNIVERSITY OF BIRMINGHAM

Department of English

Centre for English Language Studies

Open Distance Learning MA TEFL/TESL

Name	*Matthew Isaac Walsh*
Country where registered	*Japan*
Dissertation title	**The Involvement Load Hypothesis Applied to High School Learners in Japan: Measuring the Effects of "Evaluation"**
Submission date	September 21st 2009
Submission	First submission
Name of supervisor	David Field

DECLARATION

I declare:

a) that this submission is my own work;
b) that this is written in my own words; and
c) that all quotations from published or unpublished work are acknowledged with quotation marks and references to the work in question.
d) that this dissertation consists of approximately 12,750 words, excluding footnotes, references, figures, tables appendices & long quotations.

Name: Matthew Isaac Walsh

Date: September 21st 2009

The Involvement Load Hypothesis Applied to High School Learners in Japan: Measuring the Effects of "Evaluation"

By
Matthew Isaac Walsh

A dissertation submitted to the
Faculty of Arts
of the University of Birmingham
in part of fulfillment of the requirements
for the degree of
Master of Arts
in
Teaching English as a Foreign or Second Language (TEFL/TESL)

This dissertation consists of approximately 12750 words

Supervisor: David Field
Centre for English Language Studies
Department of English
University of Birmingham
Edgbaston, Birmingham B15 2TT
United Kingdom
September 2009

Abstract

This study was an attempt to test Laufer and Hustijn's Involvement Load Hypothesis concerning incidental vocabulary acquisition with learners in a Japanese high school setting. Findings suggest that with these learners learner-initiated composition with target words does lead to acquisition of them more efficiently than other output-oriented tasks such as gap-fill with context provided by the learning materials. An emerging finding is that this seems to be truer with learners that initially have a higher level of vocabulary knowledge than those with a lower level of knowledge. The reason for this could be related to the learner's ability to mentally imagine a context in English. As implication for utilizing the advantages of learner-initiated composition for vocabulary acquisition could be that providing an understandable context, or a series of them under a specific theme would provide hints for the learner concerning the field or domain of associated words and thus make it easier for them to create meaning with new words.

Contents

1. Prologue

Before the main text of this dissertation begins I would like to set out in layman's language what it aims to do and the motivation for choosing such a topic for research.

I am interested in and use TBL (Task-Based Learning), so for my dissertation I wanted to somehow investigate tasks because this would directly help me help my students. The knowledge I gained in the process of the dissertation would translate into better classroom practice if I investigated tasks.

I had been searching for an appropriate topic, an experiment that would be manageable, so I started looking for past research into tasks in EFL (English as a Foreign Language). I came across several clusters or groups of empirical research surrounding specialized topics, such as "meaning negotiation", "planning time", "task repetition", and so forth. Not being able to come up with an experiment that I felt would be manageable, grounded in an ongoing academic discussion and applicable to my teaching situation, at an academic seminar I made the brave move of approaching Rod Ellis, an authority on tasks in EFL and TBL. He pointed me to Laufer and Hulstijn's (2001a) paper describing an overarching construct that claimed to predict what sort of learning tasks would be effective for the acquisition of vocabulary by identifying elements within tasks that were either present or missing that then correlated with their effectiveness. He said I could easily create two conditions based on the construct and test the model with my learners. An interesting part of the construct for me was the claim that with other variables equal, if learners made a sentence with a new word, they were more likely to remember that word than if they had done other types of exercises like gap-fills or definition matching. This relates to my personal experience as a language learner which I will describe below, but it also relates to my perception of what sorts of learning activities my learners lack in their education.

My second language is Japanese. Throughout my career as a language learner I have tried a variety of techniques for learning vocabulary. These techniques would include things like flash cards, keeping notebooks, writing out lists, or reading simplified texts. It seemed that more often than not, new words would be forgotten. There was one technique however that did seem to be quite effective. I and a classmate from Kyoto University of Foreign Studies had created a game for learning new words. Early in the day we would choose a random word from the dictionary. According to the rules of the game we had to use the word at least once that day and then report back to each other

about it at the end of the day. I was amazed at how well this technique worked and to this day I can remember some of the words we learned in this way. They weren't necessarily common words so it seemed even more surprising that I was able to acquire them so easily. It seemed that frequency in input was not the key factor. By using it that day according to the rules of the game I often was able to remember the word, use it in output and notice it in input much later, months later. Why was I able to trick myself into remembering these words using the "make a sentence" game technique when other methods seemed to have such limited success?

In Laufer and Hulstijn's construct, the element of "make a sentence" is given the label "strong evaluation". According to their findings, having "strong evaluation" in a task; in other words having learners make sentences with new words, is advantageous and my "make a sentence" game left me with some suspicions about the nature of acquisition of vocabulary so the obvious question is will the efficacy of making sentences or "strong evaluation" hold true for my learners as well?

At the high schools in Japan I have worked at, EFL classes consist primarily of lectures by the teacher explaining items in a textbook. Chances for learners to use the language in such a way as making their own sentences or compositions are rare. Intuitively I believe that this is a very important part of language acquisition and one critical thing missing from the typical EFL curriculum in formal education in Japan. TBL can provide chances for this sort of output, but if I can empirically investigate how and to what extent for my learners here in Japan, this is an effective way of learning new words, it will certainly lead to a more informed teaching methodology for me and any other educator that happens to come across this work.

2. Introduction

This paper will summarize quantitative research comparing task characteristics in terms of factors leading to incidental vocabulary acquisition in EFL. The investigation uses Laufer and Hulstijn's (2001a) construct of Task-Induced Involvement. The participants consisted of 223 tenth grade high school learners from two different schools. The two populations were found to have quite differing levels of L2 vocabulary knowledge from the start. Two conditions were compared: a task that creates "moderate evaluation", and a task that creates "strong evaluation". "evaluation" connotes the mental process of fitting a context to a new word. This can be "moderate" if the context is contained

within the learning materials, or "strong" if the learner creates the context themselves. In both populations, the task creating a "strong evaluation" (original sentences) was found to be more effective at enabling retention one-week post task than the "moderate evaluation" (gap-fill) task. Furthermore, the population with the higher initial level of vocabulary knowledge seemed to benefit relatively more from the "strong evaluation" task. For high school learners in Japan, learner-initiated composition using L2 words new to the learner proved to be an effective way of acquiring them. A Task-based Learning approach focusing on specific content is suggested as a way to enable this output through creating context.

3. Review of the Literature

The Construct of Task-Induced Involvement (Laufer & Hulstijn 2001a) and the subsequently empirically tested Involvement Load Hypothesis (Laufer & Hulstijn 2001b) aims to categorize, operationalize, and quantify the extent and way second language pedagogical tasks focus on new vocabulary items so that incidental acquisition or retention of the vocabulary items occurs. Before explaining the framework we will define how some of the basic terms are used in the framework.

3.1 Definitions

Almost every word of the title of the article describing the framework used for this experiment: 'Incidental Vocabulary Acquisition in a Second Language: The Construct of Task-Induced Involvement'. Laufer and Hulstijn, (2001a) has a variety of meanings according to different researchers or authors.

In summary, it could be said that Laufer and Hulstijn use very broad definitions of terms such as "incidental", "acquisition", "task", or "vocabulary" presumably to allow the framework to operationally cover the widest range of pedagogical procedures or past research. In order to understand this, we must make clear how these terms are used elsewhere and define how they are used in Laufer and Hulstijn's framework and in this work.

3.1.1 "Incidental"

In a special volume entirely devoted to "incidental vocabulary acquisition" (Gass, 1999) notes that all three of the words in "*incidental vocabulary acquisition*" are

as of yet substantially unclear. She understands the term "incidental" acquisition to mean "as a by-product of *other* cognitive exercises involving comprehension." (ibid, p. 319), and quotes several other operational definitions such as "learning when the intended focus of the learning is elsewhere"(ibid, p 320), or "language learning as a by-product of language use by the teacher or anybody else in the classroom, without the linguistic structure itself being the focus of attention or target of teaching maneuvers"(ibid, p320). She however comments that these definitions bypass the active role of the learner. A focus in class is meant to pull attention to something specific, but the learners' control this focus and it is virtually impossible to measure or control, instead it is a mixture of externally driven and internally driven attention that control what is learnt. From the perspective of the learner, Gass makes the distinction between incidental and *"intentional"* learning. As is illustrated below in figure 1, difficult or unfamiliar words, if they are to be learnt, tend to be done so in an intentional way whereas frequent and familiar words might be acquired with little conscious effort. This [+/- cognate] [+/- exposure] [+/- known associated L2 words] cline could be said to relate to the Involvement Load Hypothesis in that to a certain degree, "intentional" means "involved". For example "intentional" learning implies "evaluation", the word Laufer and Hulstijn (2001a) use to mean the mental process of fitting a previously unknown word to a context. In the process of Gass' "intentional" acquisition of L2 words, collocations, and thus L2-related words would naturally be noticed. Along the same lines, a "search" as Laufer and Hulstijn use to mean the act of finding the meaning for a word, would be implied for "intentionally" learnt items that were not presented or explained within a formal educational setting to the learner.

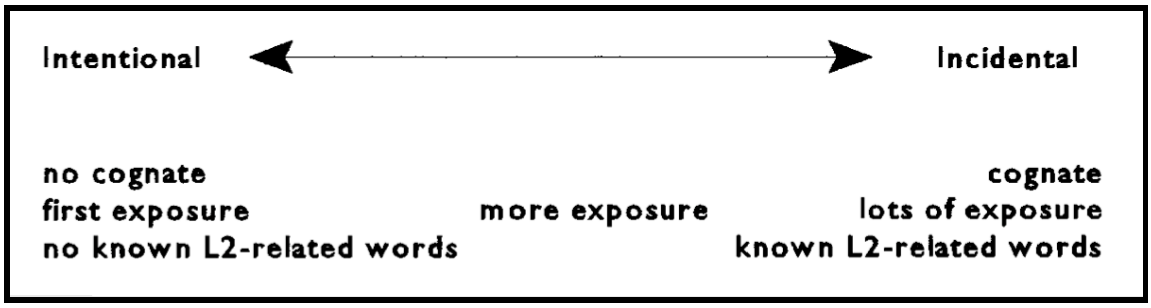

Figure 1. Intentional verses Incidental Learning

(Reproduced from Gass, 1999)

However, the definition of "incidental" in The Construct of Task-Induced Involvement

(Laufer & Hulstijn, 2001) used for the present research allows for effort to be spent learning the word, and/or the word to be the target of teaching maneuvers. The point where learning stops to be "incidental" is where it is understood by the learners that the target vocabulary item is to be part of a forthcoming test that will determine their grade. Laufer and Hulstijn are not specific as to why, but one could assume that it was because learners would be much more likely to review at home, or deploy other memorization techniques if the were to be held accountable for the knowledge of the word on a test. As described below (see "The Construct"), the operational framework aims to isolate factors of tasks in order to measure their effectiveness. Defining "incidental" like this, attempts to zero in on the task effect. In the experiment that is the subject of this paper, many such measures were taken to isolate task effect as is described below (see "Method").

3.1.2 "Acquisition"

"Acquisition" is also a term used in different ways by different authors. Ellis (1993) summarizes how in one sense it is used to mean the difference between learning and acquisition, the later equated to "picking up" a language feature. This distinction is similar to that between "intentional" and "incidental" and indeed the two sets of terms do seem to overlap in what they aim to describe. What the present research measures is described as "acquisition" because "learning" would perhaps imply the target lexis was subject to a wider array of teaching maneuvers whereas the subjects of the experiment used only one activity and were blocked from using others on their own (see "Method" below). As Ellis (ibid) mentions "setting" for the mastery of a linguistic or communicative feature of language may vary or overlap, but "acquisition" implies an attained level of competence in usage of that feature. Laufer and Hulstijn (2001a, 2001b) and the present research use "acquisition" to mean the ability to recognize and comprehend a lexical item previously unknown to the learner. It is "receptive" competence we measure. Reasons for this are further described below.

3.1.3 "Task"

In broad terms, "task" is used in two different senses; as an instrument for investigating SLA and as a pedagogical tool for organizing teaching in accordance with the premise that "if learners are to develop the competence they need to use a second language

easily and effectively outside the classroom, they need to experience how language is used as a tool for communication inside the classroom " Ellis (2003, p. ix). In other words, tasks used in research to measure an aspect of SLA, and tasks used in the classroom to create a meaning space for language use. Laufer and Hulstijn aim to investigate the nature of tasks themselves, so it is the second sense, the pedagogical task they are concerned with, however we will consider a few major authors' definition of "task" and compare them to that used for the Construct of Task-Induced Involvement.

In pedagogical research "task" is often used to mean a specific kind of activity used as a basis for Task-based Learning. Skehan (1996, p. 38) defines a task as "an activity in which meaning is primary; there is some sort of relationship to the real world; task completion has some priority; and the assessment of task performance is in terms of task outcome. Ellis (2003) very similarly differentiates task and "exercise" stating that tasks involve 1) a primary focus on meaning, 2) some kind of gap, 3) real-world processes of language (the participants select from their own linguistic resources to perform the task), and 4) a clearly defined communicative outcome. Willis (1996, p.23) states similarly that she considers tasks to be "activities where the target language is used by the learner for a communicative purpose (goal) in order to achieve and outcome". In her model "target language" to be used for the task is not initially provided by the teacher, which goes along with Breen's (1989) inclusion of "spontaneous communication of meaning" as being required by some types of tasks. Skehan's original definition of task (Skehan, 1996) is almost identical to Ellis (2003) but he later adds that "people are not given other people's meaning to regurgitate" (Skehan, 1998). These principles of the task-based approach to language pedagogy are in line with the Construct of Task-induced Involvement on many points. For example, *strong evaluation* (see "The Construct' below), linguistic output originated by the learner, can be seen in several of the definitions mentioned. Also, an activity being "communicative" could be said to create *need*; some "gaps" can create a *"search"*, etc. However, Laufer and Hulstijn's (2001a, p.16) use a broader definition of "task" stating that "…Aim(ing) to stimulate theory and empirical research (as opposed to sound pedagogical practice), it suffices to adopt (a) more general definition even though that definition encompasses artificial non-communicative tasks, such as filling in gaps or writing isolated and unconnected sentences with given words." While stating that applying the notion of involvement to the classroom would mean trying to follow the

principles of task-based language learning as defined by Skehan (1996) and the others mentioned above, they use the framework to investigate a wider range of learning activities and state they use the word "task" in the more general way, as in the Longman Dictionary of Applied Linguistics, Richards et. al. (1992, p. 373) define it "an activity of action which is carried out as the result of processing or understanding language (i.e. as a response)".

3.1.4 "Vocabulary"

Laufer and Hulstijn (2001a, 2001b) aim to investigate vocabulary acquisition, but what a "vocabulary item" means again varies in different research. Beyond single-word lexical items, multi-word items or expressions are throughout language. "Collocation" is one umbrella term used to describe these word combinations however this term again is gray. Wray and Perkins (2000) identify as many as 42 separate terms used in the literature, including "collocation" to describe what they term "formulaic sequences" and Altenburg (1990) observes that 70% of adult native speaker language may be formulaic. As Carter (1998) observes, there are in fact no set lines dividing the characteristics of these language items, rather each of them lie somewhere on a dual cline of relative fixedness and idiomacity (in Carter's terms, "opacity"). In fact, all patterns permeating language distinct from those determined solely by sentence or clause grammatical structure could be considered "vocabulary" items, including verbs or expressions that can actually determine clause structure. Although beyond the scope of this paper it is worthy to note that such observations about the behavior of lexis poses problems for both language instruction that takes a "structural" approach, assigning grammar the role of organization of parts of speech in clauses, and research on learning vocabulary that deals only with single-words. Case in point, standard word lists used in language pedagogy to narrow in on words that are common and therefore useful to the learner such as The General Service List (GSL) or the Academic Word List (AWL) are based on corpora studies of frequency and contain exclusively single-word items. These include words such as prepositions that are arbitrarily part of a fixed expression and are idiomatic in that sense. Only a very small piece of the picture is seen by either approach.

Instead of tackling these complicated issues of typology, and again to create the widest net, Laufer and Hulstijn (2001b) and this research turn to the learners themselves to ask

them which "vocabulary" is troublesome for them. This resulted in a mix of single and multi-word items. This is described in detail below in "Method". Now we will turn to the theoretical discussions The Construct of Task-induced Involvement and The Involvement Load Hypothesis aim to address.

3.2 Theory the Construct Aims to Address

The academic discussions that the Construct of Task-Induced Involvement aims to address are varied. In a broader debate concerning the role consciousness and pedagogical intervention in second language acquisition (SLA), the Involvement Load Hypothesis (Laufer & Hulstijn 2001b) is preceded by several proposed frameworks for the role of noticing, attention, and consciousness in SLA (Schmidt, 1994, 2000; Sharwood Smith, 1981; Gass, 1988; Swain, 1985; Robinson, 1995). These arguments together form what is referred to as a "weak interface" position. The "weak interface" position posits that consciousness-raising of language features occurring as a result of learning exercises can affect the acquisition process (See Figure 2 below). Cognitively, this effect would theoretically occur to the degree that explicit, procedural knowledge clarifies or corrects implicit knowledge. Explicit knowledge can be thought of as knowledge of rules of language that can be articulated. In anthropologist Edward Hall's terms, "an extension" for the reason that being able to articulate such knowledge is separate from the actual ability to apply it. In contrast to explicit knowledge, implicit knowledge is knowledge that was gained more through the learner's experience and application. It is capable of being deployed but perhaps has been left unexamined consciously and therefore in need of clarification in order to be applied with accuracy.

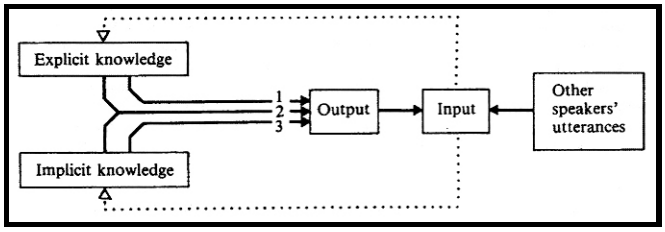

Figure 2

The Weak Interface: Sharwood Smith's model of L2 acquisition (Sharwood Smith, 1981)

This, of course is in opposition to the "no interface" position of the "Input Hypothesis

Model" (Krashen, 1982) which posits that explicit knowledge, pedagogical focus, or learnt knowledge has no impact on whether features of a language targeted in the classroom are acquired or which ones actually do end up being acquired. The "no interface" position claims that it is solely volumes of comprehensible input that unconsciously pushes acquisition forward. This process could be thought about in terms of the existence of a Universal Grammar (Chomsky, 1981), an innate knowledge of human language that is then filled in by hints about what is allowed in that particular language via comprehensible input as it directly accesses another mental tool that all humans have, the Language Acquisition Device (Chomsky, 1986). The research behind the Input Hypothesis (Dulay & Burt, 1973) shows that amount of formal learning does not influence accuracy when second language is applied in real time. From this it would appear that our consciousness does not have direct access to the LAD, and pedagogical focus or explicit knowledge are cognitively secondary to the subconscious mechanisms in control of language acquisition. This summarizes a "no interface" position on the role of learnt knowledge in the acquisition process. Below (Figure 3) is a graphical representation of the Input Hypothesis, "no interface" model.

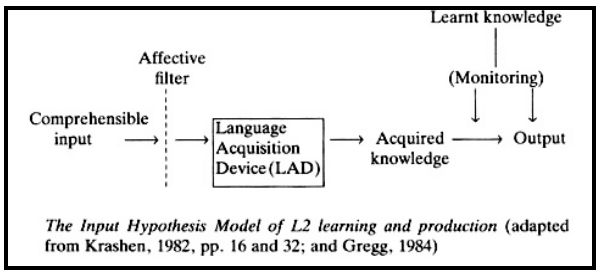

Figure 3: The Input Hypothesis Model

Concerning this larger discussion, the Involvement Load Hypothesis tacitly weighs in on the side of "weak interface". Since it is the size of the effect of pedagogical exercises on acquisition it attempts to quantify, acquisition through pedagogical focus is understood to exist in this framework. It could also be said that the quantitative research supporting the Involvement Load Hypothesis (Laufer & Hulstijn 2001b) supports a "weak interface" position by offering empirical evidence, something some of the aforementioned cognitive models of a weak interface lacked.

A problematic point here is the definition of "acquisition" as mentioned in the definitions section above (3.1.2). While the morpheme studies supporting The Input

Hypothesis (Krashen, 1982) asked subjects of the research to spontaneously produce language cued by questions that obliged the use of certain language features in order to respond, Laufer & Hulstijn's own research supporting The Involvement Load Hypothesis (Laufer & Hulstijn 2001b) asks the subjects to recall, by writing an L1 equivalent on paper, the meaning of a word unknown before appearing in a pedagogic activity done in class, at the most two weeks previously. It is obvious that cognitively there is a difference between spontaneously and correctly producing in real time language features incorporated into the learners' interlanguage over a long span of time and, recalling on paper a specific newly learned item. The question remains as to whether the effect of the initial involvement load in learning an infrequent item lasts, or whether it is negated by frequency, or lack thereof, in comprehensible input. This question is beyond the scope of the present research.

Further explaining the background of the Involvement Load Hypothesis, it is important to note that while it would seem logical to conflate the Input Hypothesis Model to include vocabulary acquisition, the actual experiments evidencing it tested for grammar usage (Dulay & Burt, 1973). Laufer and Hulstijn differ in that from the outset, they purposefully explore acquisition of vocabulary rather than grammar.

Building on a different academic discussion more about cognitive psychology than second language acquisition, the Involvement Load Hypothesis aims to examine the workings of memory and what cognitive processes might more efficiently make learned items not be forgotten. In other words, why are some things remembered and other things not. The Construct of Task-Induced Involvement sets out to operationalize the concept of depth or "Levels of Processing" (Craik and Lockhart, 1972). Craik and Lockhart (ibid) summarize how that preceding their "depth of processing" model, the prevalent "multistore" model of memory envisioned short and long-term containers for storage of information (Murdock,1967) (Atkinson and Shiffrin, 1968). In the framework, noticed or "registered", information is transferred to the "short-term store" (STS), where it then might again be transferred it to "long term store" (LTS) via things like rehearsal, repetition or association. It was believed that holding the information in STS for an amount of time sufficient enough would transfer it to LTS. Critics of this model were unsure of the idea of these memory containers being independent or demarcated, or that information must necessarily "pass through" STS as a requirement to reaching LTS (Tulving & Patterson, 1968) (Shalice & Warrington, 1970). Instead, Craik and Lockhart

(1972) propose that information is processed at different depths or levels determining different degrees of what they term “memory trace persistence”. A greater degree of semantic or cognitive analysis implies a greater depth of processing. Memory is "viewed as a continuum from the transient products of sensory analysis to the highly durable products of semantic-associative operations"(ibid. p. 676). In terms of foreign/second language pedagogic methodology, this view was the theoretical basis for “the keyword method” of learning vocabulary (Atkison, 1975). Learners created what Atkinson called an *“acoustic link”* and an *“imaginary link”* in their minds to the target word. Empirical research carried out by Atkison (ibid) and others later did demonstrate this method improved the rate of recall. For a summary see Rodriguez and Sadoski (2000). The primary concept that “memory performance is determined far more by the nature of the processing activities engaged in by the learner than it is by the intention to learn per se.” (Eysenck, 1982, p.203) is still broadly accepted by cognitive psychologists, and one can assume that “deeper is better” but an unambiguous definition of a “level of processing” has remained elusive. As Baddely (1978) concludes, “all attempts to measure processing depth appear to have been unsuccessful”. The aforementioned experiments (Rodriguez and Sadoski, 2000) compared specific techniques for memorization. However, without a construct for describing which elements within the techniques are more effective, we cannot predict which other pedagogic tasks would create more processing depth than others. We would be limited to studies comparing specific techniques.

Addressing this academic discussion Laufer and Hulstijn (2001a) attempt to quantify depth of processing as it relates to learning vocabulary by creating “The Construct of Task-Induced Involvement” (ibid) which divides the nature of processing activities created by a task into three components, *“need”, “search”, and “evaluation”*. A pedagogic task containing more of the components would produce better retention of the vocabulary items. This is explained in more detail below. The components themselves were abstracted from elements found in other effectiveness studies. (for examples, see Table 1 below)

3.3 The Construct of Task-Induced Involvement

Of the three components mentioned above, “*need*” is actually more of a driver than a “processing activity”. It is the *motivational* component of the construct and simply

means the need to achieve or drive to comply with the task requirements. In the case of vocabulary items, this translates into the need to understand or perhaps produce the word in order to complete the task. "*search*" is one of the *cognitive* components of the construct and means the process of finding the meaning of unknown L2 word, or finding an L2 word expressing a certain concept. In second language pedagogical tasks this would typically translate into referring to a dictionary, or negotiating meaning with a pair-work partner. The last component "*evaluation*" is the cognitive process of comparing a given word with the surrounding words, meanings, or contexts. This could mean any sort of gap-fill exercise where the surrounding words, meanings, or contexts obligate a certain choice of words, or in more of a productive sense, where the learner must provide other words, meanings, and contexts that match the word to be used. "*Evaluation*" is the process of making a selective decision considering the form, usage, collocations, and meaning in general of a certain word.

Part of the construct is also a moderate/strong distinction in the case of *need* and *evaluation*. A *search* is something that either is done or not done but in the case of need, there is a difference in involvement if the learner initiates the need for a word; as in if the learner had a concept that they wished to express as part of an original composition rather than if the task had simply asked them to use a given word. This is also true in the case of evaluation. Recognizing differences between words such as in a fill-in task is less "involving" than making a sentence with one and choosing additional words to combine with the target in use. The distinction between moderate and strong *need* or *evaluation* is that of externally provided verses learner initiated opportunities for use. Examining this distinction further on a theoretical level certainly would be interesting but a bottom-up examination based on the empirical data of past experimental research also reveals the durability of the construct of task-induced involvement load including of course the moderate/strong distinction. We can see in this in the table below.

The more effective task	The less effective task	Reference
Meaning selected from several options +evaluation	Meaning explained by synonym	Hulstijn 1992
Meaning looked up in a dictionary +search	Reading with/without guessing +/–search	Knight 1994; Luppescu and Day 1993
Meaning looked up in a dictionary +search	Meaning provided in a marginal gloss	Hulstijn *et al.* 1996
Meaning negotiated ++ need, +search	Meaning not negotiated	Newton 1995
Negotiated input +search	Premodified input	R. Ellis *et al.* 1994
Used in original sentences ++evaluation	Used in non-original sentences	Joe 1995, 1998
Used in a composition (L1–L2 look up) ++evaluation	Encountered in a reading task (L2–L1 look up) –/+evaluation	Hulstijn and Trompetter 1998
Interactionally modified output ++evaluation	Interactionally modified input	R. Ellis and He 1999
Reading and a series of vocabulary exercises +evaluation/++evaluation	Reading only (and inferring meaning) –/+evaluation	Paribakht and Wesche 1997
Reading, words looked up in a dictionary +search	Reading only, words not looked up	Cho and Krashen 1994

Table 1: Previous research in terms of Involvement Load

(from Laufer and Hulstijn, 2001a)

*-/+ indicates that the component was only presents for select parts within the task.

We can then see in the table below (table 2) of the scoring system that the plusses (+) indicate the degree of involvement in any of the components and the total amount of plusses indicates the total involvement load of a given task.

Involvement Load

	Need	Search	Evaluation
none	( )	present (+) or not ( )	( )
moderate	(+)		(+)
strong	(+ +)		(+ +)

Table 2: Scoring of Task Involvement Load

4. Research Questions:

The present research compares the weight of moderate and strong evaluation. The research design is described in detail below but we start with the questions:

1) Similar research has been conducted in Israel and Amsterdam (Laufer & Hulstijn, 2001b) with higher proficiency level learners. In Japan, at the high-school level, with learners who have a much less vocabulary knowledge, and where such open-ended productive activities are traditionally rare in formal education, will "learner-initiated" composition (strong evaluation) tasks also be found to be more effective for vocabulary acquisition compared to activities where learning materials provide context and learners are asked to match those contexts with a new word (moderate evaluation)?
2) What are the implications of the findings to classroom practice?

5. Method

5.1 The Tasks

Both versions of the worksheet are attached as appendix 1. Both tasks accompany a 326-word text studied as part of the normal syllabus. Entitled "Child Labor", the text was written by the instructor/researcher with EFL pedagogical purposes in mind and was made to be within reach of the learners in terms of complexity and vocabulary. Both versions of the worksheet involve the text first being read for comprehension. The comprehension questions at the end of the texts are engineered so that the target vocabulary must be understood in order to answer them. This creates a "need", the motivational component of Laufer and Hulstijn's (2001a) construct. In this case, the comprehension questions create a "moderate need" because the external agent of the task imposes the need. Both versions of the worksheet were also identical in that they did not invoke the component "search" (Laufer & Hulstijn, 2001a). One of the tasks, "Task A", does however involve searching for a word but not in the involvement load sense. Blanked out spaces in the text are matched to words with L1 glosses arranged in a jumbled matrix. According to Laufer, (personal communication, February 11, 2008) this would not be considered a "search" since the learner does not need to hold the

unknown word in memory (in the phonological loop) long enough to, for example, consult a dictionary. In other words, there is not enough effort in simply scanning through the matrix. Task A however does create Laufer and Hulstijn's cognitive component of involvement, “evaluation”, because the learners' goal is to consider or “evaluate” whether target words and sentence context and/or language match. In this case it is a “moderate” evaluation because the contexts and surrounding language are externally provided or given. If the learner were creating original sentences with new vocabulary, the level of evaluation would be considered “strong” according to the framework. This was the case with Task B. In Task B target vocabulary items were not gapped out in the text as in Task A but underlined and on the side there was a list, in order, of the words again with L1 glosses and a blank line for the learners to compose an original sentence using the vocabulary item.

In summary, both versions of the worksheet have “need” from the comprehension questions, neither has “search”, and the varying condition is between the “moderate” evaluation of the matching gap-fill Task A and the “strong” evaluation of original sentence Task B. I will next describe the environment and participants in the experiment, choices in the process of creating it, and considerations in its execution.

5.2 Populations and Groups

As is illustrated below in Table 3, the experiment described in this paper was conducted identically with two separate populations consisting of three groups each. The populations both consisted of tenth grade high school students in Osaka, Japan but from different schools with different levels of academic achievement or aptitude as determined by entrance requirements and the resulting stratification of the educational system as is described below. These populations will be referred to as “School A” and “School B”. There were three classes, or “groups” from each school consisting of between 35 and 41 learners totaling 223 participants, but when accounting for missing data due to people who were absent for either the pre or post-test, the numbers in the groups lessen to between 30 and 37 with a total of 203 participates in the experiment.

Population (school)	Group (class)	Number of participants in group after missing data	Role in experiment
School A	10th grade class 3	37	Control (No Task)
School A	10th grade class 4	38	Task A (Gap Fill)
School A	10th grade class 2	37	Task B (Orig Sentence)
School B	10th grade class 14	30	Control (No Task)
School B	10th grade class 16	30	Task A (Gap Fill)
School B	10th grade class 15	31	Task B (Orig Sentence)
		total=203	

Table 3: Populations and Groups.

The two population's differences in terms of English vocabulary knowledge are described later, and a description of the schools' backgrounds will explain the general differences in environment and peers. In Japan, a high school's standard or ranking varies in relation to the entrance requirements imposed by the institution, including of course the school's entrance exam. This "level" then usually correlates with the percentage of graduates continuing on to university, and then to the quality of the universities they continue on to. National universities are considered more prestigious in general than private universities of which however, a select few are well known and also considered prestigious. Returning to high schools, public high schools are known to be at both extremes, the best and the worst, while private schools occupy the middle range. The ratio of public to private high school students enrolled is approximately 7 to 3 respectively. School "A", the more prestigious of the two, is public and is in fact one of the very few national high schools in Japan and is attached to a national university of education. Its entrance exam is considered to be the most difficult in western Japan and a relatively very high ratio of graduates continue on to national universities. School "B" the lower academically ranking of the two is a private high school, of which there are many in Japan, and is associated with the Episcopal Diocese of Osaka. Its graduates

generally continue on to private universities, often to the well-known ones. In terms of entrance exam difficulty as well, the general quality of the institution is considered better than average. When comparing these two populations it would not be accurate to say that School B consists of particularly low aptitude learners, but rather that School A could almost be considered a special school for gifted learners.

In order to verify the difference between the populations specifically in terms of knowledge of vocabulary in English, Nation"s (1990, pp. 265-266) 2,000-word level test was used. The multiple-choice test was designed so that the percentage of correct answers out of a possible 18 correlates to the learner's vocabulary knowledge of the 2000 most frequent content words of English. For example, a score of 9 (50%) would mean a learner would understand about 1000 of the most frequent 2000 words. Combining the scores of all 3 groups from each population, School A's average score was 13.16 (73%), or a presumed knowledge of 1462 words whereas School B's average was much lower, 8.07 (45%), or a presumed knowledge of only 897 words. Further investigation of the root causes for this sharp difference, whether along the lines of socioeconomics or a result or self-perpetuation of the described above stratification of the educational system go beyond the focus of this study. However, the differences as seen in the vocabulary level tests do reveal some interesting patterns as will be explained in the results section.

5.3 Choice of Lexical Items to be Tested.

The object of the experiment was to test the effect of the tasks on retention of the vocabulary items, but the items would first need to be deemed likely to be unknown to the test subjects if a gain would result that we could measure.

A common technique for this used in the research is to target words from the Academic Word List (AWL) (Laufer & Hulstijn, 2001b) or from the upper bands of frequency lists (Webb, 2005). Relatively rare words are assumed not to be known by the majority of the experiment population. Issues involved with targeting words from lists like the AWL will be discussed presently.

Webb (ibid) goes even further than this and substitutes the list words for nonsense words to completely ensure they are unknown at the onset. This technique also eliminates the need for a pre-test and the danger of a learning effect from it. However this method would not have fit our purposes since the target words were to be found in

reading texts that existed as part of a course that was already in progress. Aside from the obvious difficulty of asking class participants to waste their time learning nonsense words, the use of such words seemed a little unnatural and problematic to this researcher for other reasons as well. For example, in Webb's research the relatively low frequency word “locomotive” is replaced by the nonsense word “masco”. It could be said that “masco” does not sound like what it is supposed to mean, or rather it sounds more like something else. Pinker (1994) notes that although the relation between a word’s sound and its meaning is arbitrary, factors like onomatopoeia and “sound symbolism” can play a role. As an example of sound symbolism or the sound of a word calling to mind what it refers to, Pinker (2008) mentions the words “bling” meaning ostentatious jewelry or “blog” which, sounding like “blurb” or “blob” does fit well the image of a mass entity of informal writing on the internet. Pinker (1994) also mentions “phonetic symbolism”, in specific vowel sounds. Higher frequencies (sounds, tones) can remind people of smaller things and visa-versa. He gives the example of “bit” in English, or the Chinese words for “small” and “large” being ch’ing and ch’ung respectively. When given the choice, native speakers of English can often correctly guess which Chinese word means which. This writer informally asked a few native speakers of English to guess what “masco” meant and the response that it “sounded like” the name of a company was received repetitively. Perhaps the “co” brought to mind companies like Tesco or Costco. In any case, this also supports the idea that words do tend to “sound like” something. Crafting nonsense words to be neutral in these respects would certainly not be easy.

Another issue might be the level of suspension of disbelief required to accept “masco” as the English equivalent of “locomotive”, a word virtually synonymous with “train” which is a 500-1000 level word in English and very salient in the learners' mental lexicon in Japan, where the research takes place.

In other words, one could easily imagine that at some affective or psychological level the learner would not initially accept “masco” as “locomotive” or “train” and therefore would not be able to recall it at post-test. In a related sense, had the researcher simply used the real word “locomotive” and the learner had accepted it, the morphemes or constituents making up the word perhaps might have given meaning related hints to the learner and created associations aiding in retention. Possibly also the learner would have had scant exposure to the real word in the past, but not necessarily be able to fully or

accurately recognize or produce it but be able to infer or estimate correctly its meaning from this vague memory of contact with the word when in the present environment. It is assumable that Webb wanted to avoid these sorts of aids but I propose that they would not necessarily lessen the validity of the results and in fact having such factors present is arguably closer to the cognitive processes at work during a real learning situation. Furthermore, this gets to the question of what we are trying to test. To the extent that tasks act as consciousness-raising activities, partially known words being pulled closer to mastery is certainly an aim of tasks and the ability of tasks to achieve this certainly constitutes their effectiveness. In summary, Webb's use of nonsense words as targets items, beyond being literally contrived to a problematic level, only measures the effectiveness of the tasks bringing words from the outermost rim of consciousness closer to mastery. We felt, for reasons outlined later, that receptively the ability to produce an L1 equivalent of the target would be the simplest and most prominent gain to measure. We therefore decided to allow for the possibility of previous contact and authentic words were used as target items in this research. However, words to target were selected carefully.

As mentioned before, the words need to be generally unknown to the populations in order for there to be a significant difference in pre and post-test results and between treatments in order to attempt to attach causality to the varied factors of the tasks. Selecting authentic words found in frequency lists such as the AWL is a common way of choosing words to be targeted in similar research but the way the words were chosen in this experiment turned out to be significantly more efficient at teasing out words unfamiliar to these specific populations. I will offer a comparison and analysis of target words had they been chosen using frequency lists opposed to the way it was actually done, but first I will describe the method used.

In order to identify words from the text likely to be of difficulty to the groups, the experimental populations' peers and seniors were asked to assist. Students not taking part in the experiment but from the same schools, a total of 28 spread evenly between the 10th, 11th, and 12th, grade, were given the text and asked to read it and underline as many unfamiliar words as they could find. This was done with non-participating students so as not to affect the results by exposing participants to the text before the experiment. Also asking students in all 3 grades of the high schools to cooperate in identifying unknown words was intended as a way of identifying the particularly

difficult words for these populations. Presumably upper level students would underline fewer words than lower level students although the lower level students would often underline the same words as the upper level students. The particularly difficult words would be underlined across all grades and therefore be underlined more often in total, whereas the less difficult words would already be known by the participants in the upper grades and thus underlined less; this lessening the total. This turned out to be true. From the 28 students, a total of 33 words were underlined and several of them were underlined by as many of as 16 of the students spanning all three of the grades. The words chose to be targeted were the 15 most commonly underlined words. The words can be seen in appendix 2.

In order to compare these results to those that would have been found by using word-lists, we used Cobb's (2008) online vocabulary profiler, which is a version of Heatly and Nation's computer program “Range” (1994). It sorts any text into four categories: K1 (the first thousand most frequent words in English, K2 (1001-2000), words from the AWL (all of which are beyond the 2000 level), and “Off-List” words, which are beyond the first 2000 and the AWL combined. As is illustrated in the table below, if we had chosen words in the text to test for by choosing those on the AWL and the upper-band K2 (1001-2000) level lists, for the most part the same words would have been found. Significant differences do however exist as can be seen in Table X below. Looking first at the word-list side we can see that from the K2 band and the AWL, the populations had not identified the words (labeled with asterisks) “coffee", “international” or “gap” as problematic. This can be accounted for by the fact that all are popular loan-words commonly used in Japanese. One exception however “labor”, from the AWL was part of the title of the text “Child Labor” and therefore probably inferred or otherwise understood from the outset.

One would expect “off-list” words to be more difficult because they are beyond K1, K2, or AWL in terms of frequency, however there were several off-list words not peer-identified as difficult. As can be seen below, these include “soccer”, “banana”, and “coconut”, “Japan” and “Nike” are proper nouns and therefore excluded from the lists but again variations in the host culture of the texts or corpora that the lists were derived from could be seen as a factor. More so perhaps, especially in terms of the Off-List words, a simple non-correlation between salience and frequency lists could be responsible. As Nation (1990, p.20) notes; "The most serious problem with

word-frequency lists is that certain useful and important words do not occur in the first or second thousand words.... "

Looking next at which words were peer-selected as unknown (Table 4 below) we can see in the opposite sense that again, the predictive value of the lists is not completely accurate. We may have chosen to target K2 and the AWL since we can see that most of them found in the text were those most identified by the peers, however when comparing totals, the amount of most identified words from Off-List (4) matches the number of words from AWL (4), whereas most of the words in the text from Off-List were not selected by the peers. As well, a few presumably low level words from K1(2) were among those most often identified as difficult.

We can see from this analysis that when trying to predict which words in a text are difficult for learners, for research or pedagogic purposes, low frequency does not necessarily equate being unknown nor does high frequency mean being unproblematic for a specific individual or population. For these populations targeting K2 and the AWL would have produced only about 56% of what they actually claimed as unknown and targeted about 31% words already known to the majority. Ascertaining the problem words for these specific populations by surveying peers produced a more accurate set of data to work with.

Top peer selected as unknown	*(out of 28 students)*	*Status of peer selected word in 'vocabprofiler' lists*	*All content words in text on 'vocabprofiler lists minus K1 (0-1000)*
gymnasium	(16)	Off-list	*From AWL*
agriculture	(16)	K2 (1001-2000)	aware
(cash)crops	(15)	K2 (1001-2000)	demonstrations
to be aware	(15)	AWL	globalization
blame	(15)	K2 (1001-2000)	labor *
demonstration	(14)	AWL	*From K2(1001-2000)*
fiber	(13)	Off-list	agriculture
increase	(12)	K1 (1-1000)	blame
plantation	(10)	Off-list	clothing
warmth	(8)	K2 (1001-2000)	coffee *
globalization	(8)	AWL	crop
sweatshop	(8)	AWL	gap *
partly	(8)	K1 (1-1000)	international *
shrimp	(7)	Off-list	typical
clothing	(6)	K2 (1001-2000)	warmth
typical	(6)	K2 (1001-2000)	*From 'Off-List'*
			bananas *
			cash
			coconut *
			fiber
			gymnasium
			Japan *
			Nike *
			plantation
			shrimp
			soccer *
			* = Not peer selected as unknown

Table 4: Wordlists verses Peer Selected Target Vocabulary

5.4 Pre and Post-tests

The pre and posttests of knowledge of the targeted vocabulary items were receptive tests. The translation test asks for an L1 (Japanese) equivalent of the target words. There were fifteen words targeted. The vocabulary items were highlighted within an example sentence and to the right, a blank was left for the learner to write the Japanese word that correlated to the target. The layout of the pre and post-tests varied slightly. Both can be seen in appendix 3.

One point was given on an all-or-nothing basis for each target word correctly translated. This varies from more complicated ways of measuring degrees of vocabulary knowledge in similar experiments. Folse's (2006) adaption of Paribakht & Wesche's (1997) method verifies not only receptive knowledge via requesting an L1 correlation in a very similar way to what was done in our experiment, but also tests productive knowledge asking for the testee to create an original sentence using the target item. This

method possibly has the advantage of enabling the researcher to measure gains higher on the scale of mastery, such as correct usage or collocation, but original sentences would have been impractically time consuming for the experimental groups at hand. In fact Task B, which involved creating original sentences, took the whole class period whereas the pre and post-tests needed to be done as a short side-activity after the lesson from the real syllabus. It was also felt that if measurable effects were to occur as a result of the tasks, they would occur prominently at the first, receptive level since from the outset the targets had been determined largely unknown to the populations. Flynn's (2007) checklist receptive test allows for lower level partial knowledge of a word, for example, "I'm pretty sure I know that word". This also arguably enables measurement of more subtle gains in vocabulary knowledge (Anderson and Freebody, 1983 in Nagy, Herman and Anderson, 1985) but it also creates the need to estimate and subtract a group's rate of false claims when given this freedom. Such self-assessment testing methods would also be necessary where a multi-linguistic learner L1 environment existed or where otherwise bilingual marking could not be accomplished. The reason being that without raters capable of marking answers in the learners' L1, English would have to be used by learners to answer the questions. The language necessary for a correct answer could easily end up creating a hurdle higher than the target words themselves. Laufer (2008b) comments that such techniques for measuring degrees of knowledge can result in scoring that is a mess. For these reasons, and because of the fact that we had no problem with scoring translations, it was decided that an L1 translation test would be neater in scoring and results and perhaps be more objective.

During the scoring of the pre-and post-tests three raters were present, one native English speaker proficient in Japanese and two native Japanese speakers proficient in English. When questions arose concerning which Japanese words were to be considered valid as equivalents to the targets, a consensus was reached and adhered to throughout the marking process, marking recursively when necessarily. Laufer (ibid) suggested giving half a point for semantically close but not exact translations but there was no occasion where this was necessary. Japanese and English words referring to the same things often vary in semantic range or usage. Because of this inherent fuzziness, all or nothing judgments seemed rather cut and dry. The main criterion was decided to be whether or not the rater could imagine a way to say the express the meaning of example sentence in Japanese with the word the learner had written as an answer.

Changing topic to the actual execution of the pre and post-tests, the time allotted for completion was the same for pre and post-test, about 15 minutes, although almost all participants finished much earlier.

Several efforts were made to prevent or minimize any learning effects that could possibly have resulted from the pre-test. The danger of the pre-test pulling attention to these words in the learners' consciousness to the extent that they might be recognized in the tasks and affect the results of the post-test was clear, and the results of the no-task control group showed that this may have happened to a very slight extent but it did not reach statistical significance, (see results section below) Several measures were taken to lessen the chance of this happening.

Firstly, in terms of timing, the pre-test was done two weeks before the tasks in hopes that any words would be forgotten whereas the post-test was done only one week following the tasks. Testing one week after task is an interval commonly used in similar research because after that, any conscious-raising effects of task factors on memory dwindles. So conversely, waiting two weeks after a pre-test that was not designed in any way to promote retention would most probably leave us absent of a learning effect for the task 2 weeks later or the post test the following week. The 3 week delayed effect of pretest to post test in control (no task) groups was found to be minuscule and nowhere near significant in the data.

Next, the example sentences in the pre and posttests containing the target words were made so that the context or collocations would not provide overt hints to the meaning of the target word, but on the other hand would not be so stripped of context as to seem preposterously unnatural. The criterion for the level of contextual clues acceptable was simply this researcher's intuition. These example sentences were of course different from the sentences in the task texts, although target vocabulary items were used in the same sense.

Then, the order of the example sentences on the test sheet was changed between the pre and post-tests to avoid any chance that the sequence could somehow trigger memory of the three-week earlier pre-test and provide some hint. It was hoped that the post-test would appear to be an altogether different test. Changing the wording of the sentences themselves between pre and post-tests was considered but it was felt that to the slight extent that guessing was possible from the context of the example sentences, changing the sentences would have the undesirable effect of creating another variable, it became a

matter of which sentence provided an easier context to guess from. Again, we can infer from the control group data that leaving the sentences the same did not lead to any significant learning effect.

Next, it was announced that the test and the vocabulary items were completely unrelated to the course, would not be tested on again, and were merely part of small research project by the instructor. Learners should not be concerned if they did not know the meaning of the words and would not be held responsible for them later. This announcement needed to be made not only to lessen the chance of a learning effect, but also because in terms of definition, learned or acquired items not being part of a test for class credit is what is meant by "incidental learning" when speaking of Involvement Load.

Along the same lines the researcher made a similar announcement for the tasks, saying that the text "Child Labor" would be tested for general comprehension as part of the final exam for the semester, but the exercises (tasks) or words they targeted were not directly relevant to the test at all.

It was also purposefully not mentioned that the words from the pre-test were same as the ones targeted in the tasks. Perhaps because of the two week interval between pre-test and task, of the 223 learners comprising the populations only 2 vocalized that they had recognized the words. When they had done so, it was not in a way that alerted the surrounding students and no other learners in fact did react.

Finally, in a further effort to zero in on the effects of the tasks rather than other individual or social factors both the pre-tests and then the task worksheets with the target items were collected immediately after completion in order to prevent the more diligent students from looking up the words later. As the sheets were being collected, students were asked not to talk or discuss the problems with each other. This was important since it was known by the instructor/researcher that in each population (school), there were clear differences in levels of talkativeness and cooperation between groups (classes). It is easy to imagine that one group's ability, or lack thereof, to spontaneously create a collaborative social space or Zone of Proximal Development (Vygotsky, 1934) aimed at making sense of the target words could easily have outweighed the effect of the factors of the tasks. Had we allowed even a short ad-lib peer feedback session to ensue as we were collecting the papers, the results surely would have been ruined since the more talkative and constructive classes would be left

with a memory to associate with the targets. The effects of such spontaneous collaboration could certainly be the subject of further investigation and one could even imagine need, search, and evaluation fitting into a framework to analyze such peer exchanges. Next I will present an analysis of the actual results of the experiment.

	Control Group (No-Task)	Moderate Evaluation Group (Gap-Fill)	Strong Evaluation Group (Original Sentences)
School A (Vocabulary level test score 1462)	Pre 5.93 (40%) Post 6.43 (43%) *Change +0.5 (3%)*	Pre 6.18 (41%) Post 10.11 (67%) *Change +3.93 (26%)*	Pre 6.89 (46%) Post 11.81 (79%) *Change +4.92 (33%)*
School B (Vocabulary level test score 897)	Pre 2.52(17%) Post 2.7 (18%) *Change +0.18 (1%)*	Pre 1.9 (13%) Post 6.24 (42%) *Change +4.34 (29%)*	Pre 2.14 (14%) Post 6.79 (45%) *Change +4.65 (31%)*
* Pre and Post indicate group mean scores are out of a total of 15 possible correct answers.			
* Vocabulary level indicates mean score on Nation's 2000 word vocabulary test.			

Comparing mean change of scores on pre to post-tests, how much more change did original sentences create over gap fill for each population or school? (Strong vs. Moderate Evaluation)	School A (Vocabulary level 1462) = +0.99 (7%) P=0.01 *(statistically significant)*	School B (Vocabulary level 897) = +0.31 (2%) P=0.39 *(not statistically significant)*

Table 5: Results of Study

6. Results and Analysis

Before an analysis can be understood, the numbers and terms in Table 5 above need to be explained or reviewed. School A and School B represent different populations. The differences of these populations were explained before, but most relevant is the difference in knowledge of vocabulary in English. Testing the populations for approximate number of words known in English resulted in the numbers that can be

seen on the table: 1462 for School A, and 897 for School B. These numbers represent the mean score on a test that estimates of the number of words out of the first 2000 most common words in English that the learners in each population know. It can be seen that School A has an initially higher level of knowledge of vocabulary. This can also be seen when comparing the pre-test scores for all groups: control group (no task), gap-fill group (moderate evaluation), and original sentence group (strong evaluation). Consistently School A scores higher than School B even before any classroom work on the target words takes place. The scores labeled "Pre" and "Post" represent the mean score out of a possible 15 correct answers on pre and post-test, as explained before, the 15 words from the text determined to be most likely difficult for learners in the populations. The percentages in parenthesis represent the mean score divided by 15, the total number of possible correct answers. The percentages of change are also calculated this way.

"Change" represents the difference between the mean scores for the groups from pre-test to post-test. In this experiment it happened to be that in all cases *"change"* turns out to be improvement from pre to post test but as is explained later, with the control groups from both populations (schools), the change did not reach statistical significance. This indicates that there was little or no learning effect from the pre-test. This can be assumed to be a result of the care taken to prevent this as described in the methods (5.4) section above. The lower part of the table illustrates the difference in change or gains when comparing the effects of the gap-fill (moderate evaluation) to the original sentences (strong evaluation) exercise for each population. Percentages again are out of 15 possible correct answers but also displayed are the t-test P value scores. It was felt that whereas with the other results, statistical significance was rather obvious and could be inferred intuitively further illustration was needed with the comparison of gap-fill (moderate evaluation) and original sentences (strong evaluation) in terms of change they created

Moving forward and interpreting the findings, we can see a substantial change representing improvement in scores resulting from both types of exercises. Multiple t-tests also verified that both conditions, gap-fill (moderate evaluation), and original sentences (strong evaluation) for both populations (School A, and School B) resulted in changes in scores that reach statistical significance, defined as $P<0.05$ as is common in this type of research. The T-tests also verified that the slight improvements we see in

the control groups between pre and post-test were not statistically significant. The t-tests that verify significance for change created by both gap-fill (moderate evaluation) and original sentences (strong evaluation) by comparing pre to post-tests (appendix 3) within the groups were single-tailed, paired (dependent) t-tests. T-tests comparing change created by gap-fill (moderate evaluation) to change created by original sentences (strong evaluation) within populations (schools) were single-tailed unpaired (independent) t-tests. For a list of the specific t-tests and their results, see appendix 4.
An interesting emerging finding can be seen when comparing the gains from the gap-fill (moderate evaluation) exercise to those from the original sentences (strong evaluation) exercise across populations. Although original sentences (strong evaluation) results in more improvement than gap-fill (moderate evaluation) in both populations, there is a difference as to the extent that is does. School B, the population with the lower initial vocabulary score according to Nation's 2000 word vocabulary test (Nation, 1990) gained relatively less from doing the original sentences exercise over the gap-fill exercise than did the population with the higher initial vocabulary knowledge (School A). In fact, for School B the difference does not reach statistical significance. A t-test gives us a p-value of 0.39 and the null hypothesis cannot be rejected. In this study for the population with a lower level of knowledge of vocabulary, making original sentences seems not to have had that much advantage over gap-fill exercises, but for the higher vocabulary knowledge population the difference was drastic. Learners who have attained a certain level of vocabulary have much more to gain from testing out new words by creating sentences and contexts with them for themselves than by looking at them in a context created by an external source. For learners of this vocabulary level, making or "evaluating" their own contexts for a new L2 word makes it more memorable. We can assume there is a deeper level of processing that leaves a higher level of "memory trace persistence".
The results of this experiment suggest there is a difference in the "processing depth", or effect of the previously mentioned types of activities according to the learner's level of vocabulary knowledge, but does not offer clues as to why. A number of possible explanations for this difference could be explored. Both of these activities involve a context that needs to match the new word. If context is the key, perhaps for the population with more limited vocabulary knowledge, the original sentences were not much better at creating a context for the unknown word than the gap-fill exercise. The

definition of “strong evaluation” includes that the learner creates the context to use the new word in, whereas “moderate evaluation” connotes fitting the new word into a given context as part of the task. For a learner with more limited vocabulary knowledge, the context provided in the reading text may be less understood than the L1 gloss provided in the worksheet. Such a learner would be tempted to pay less attention to the context provided by the text and create an original sentence depending greatly if not solely on the L1 gloss. Along the same lines, a learner with a lower level of vocabulary would have less of a L2 lexicon to draw from while creating a context for the original sentence. The original sentence could easily end up a lexically sparse transliteration of a context one would more likely find the L1 equivalent in.

The previous explanation offers a plausible reason why original sentences may not have been as effective as expected for learners with a low level of vocabulary. The next explanation will offer a plausible reason why perhaps the gap-fill was more effective than expected. We can see from the table that although the gap-fill was not actually more effective than the original sentences, it was very close. The gap found with the higher vocabulary knowledge group was missing. This explanation also centers on context.

Although both types of tasks provided an L1 gloss, the gap-fill may have had low-level learners attention more focused on the context within the text. Looking at the text-provided contexts could not be avoided. This process of considering the new word’s context, usage and collocation may have created a deeper level of processing for a learner compared to one who skipped that process as was possible with the original sentences task. The higher vocabulary level population perhaps did not skip the process even with the original sentences task, since they would not be overloaded by doing both; considering the context within the text, and creating their own context.

The question remains as to whether for the lower proficiency level population the gap-fill was more effective or the original sentences were less effective.

Even if not reaching statistical significance, in terms of mean scores, the original sentence task did outperform the gap-fill task for the population with the lower level of vocabulary knowledge as well. This perhaps demonstrates the strength of learner-created contexts for vocabulary acquisition even for the population with the lower level of vocabulary knowledge, who might have been at a disadvantage to create original sentences or understand the contexts in the L2 text.

Turning to the population with the higher vocabulary knowledge level, to explain the drastic difference in acquisition of target items between the groups, as mentioned before, simply the opposite may have been true. That is to say, with a higher level of vocabulary knowledge this population is able to tease contexts, usage, and collocations from input (the text) and they would also have greater freedom to create contexts that fit with the new word utilizing their much larger L2 lexicon. In the literature review, the keyword method was explained. This method aimed to create deeper processing by creating mental associations and purposefully connecting them to the new word. Although artificial, it creates a context containing associations. For the higher vocabulary group it may be easier to create associations with other L2 words. This would be true in a task with a text-provided context such as the gap-fill, but evidently even truer in a task where strong evaluation or original context is a component. Laufer and Hulstijn (2001b) also mention other research showing words used in productive tasks, particularly in original contexts were remembered better than those practiced in non-productive tasks (Ellis & He, 1999; Hulstijn & Trompetter, 1998; Joe, 1995, 1998). It seems that the population with the higher vocabulary knowledge was at a double advantage, able to understand text-provided contexts and able to create their own. With this footing, they were able to clearly demonstrate the cognitive advantages of creating original sentences.

In summary, the answer to the first research question is yes, strong evaluation proves more effective than moderate evaluation with high school learners in Japan whose level of proficiency is much lower than the learners in Laufer and Hulstijn's (2001b) research. However, this seems to be truer for learners with a higher level of knowledge of vocabulary to start out with. A very interesting area for further research would be investigating precisely at what levels of vocabulary knowledge do certain techniques become more or less advantageous. These findings have implications for language pedagogy, which will be explained below.

7. Implications

To answer the second research question, the findings suggest that for the typical high school student in Japan learner-initiated composition tasks with new vocabulary items is efficient for acquisition of them. Another researcher, Martinez-Fernandez (2008) in her recent in-depth empirical examination of the Construct of Task-Induced Involvement

finds that output-oriented tasks were more effective generally than input-oriented tasks for retention of vocabulary. This study suggests that learner-initiated composition is more effective than other output-oriented tasks such as the gap-fill, reinforcing the claim that strong evaluation, or learner-initiated output is key to vocabulary acquisition. Linked to this, it appears that context is important in both input and output. The task for the educator then is how to construct tasks that take advantage of these observations. A framework is needed that facilitates output and at the same time pulls attention to context in input so that new words will be noticed. Utilizing the Construct of Task-Induced Involvement would entail seeking to create as many of the conditions as possible, noting that according to the findings of this study, output, in specific learner-initiated composition involving target vocabulary is definitely an advantageous, if not the most advantageous element of the construct to have in place.

Two things would address these aims: First, the tasks should be set in a framework that facilitates learner-initiated output such as Task-based Learning (TBL), and second, that the tasks share a theme, topic or a series of interlinked topics as in approaches like CBI (Content-based instruction) (Brinton, 2003) or ESP (English for Specific Purposes). The use of a theme, increases the possibility that the same vocabulary items will appear repetitively in a series of tasks and contexts. Snow et al. (1989) label this as "content-obligatory language" and "content-compatible language" in their "Conceptual Framework for the Integration of Language and Content in Second/Foreign Language Instruction" stating "in real life people use language to talk about what they know and what they want to know more about, not to talk about language itself" (ibid, p. 202). Along the same lines Snow et al. (1998, p.202) quote Mohan (1986) "In subject matter learning, we overlook the role of language as a medium of learning. In language learning, we overlook the fact that content is being communicated." To solve the problem of integration of language and content teaching, Ellis (2003) proposes a modular approach whereby the syllabus is dominated by unfocused tasks (no particular linguistic items specified to be used) around a certain content area. Later, or as learners advance, focus on form is used in a remedial way. In terms of context, Ellis (2003) describes Cummings' (1983) model of language proficiency and how a task embedded in a context known to the learner creates less of a cognitive load than a task that is not supported by a known context. This would imply that for learning unknown, infrequent, or otherwise difficult vocabulary items a familiar context will be of aid. On the other

side of the coin Ellis (2003 p.95) states that "…cognitively challenging tasks…may promote acquisition" quoting research about interaction and how difficulty creates opportunity to use L2 for communication and thus pushes acquisition forward. Ellis (2003 p. 95) adds the caveat that " if a task is too challenging it may simply cause learners to give up!" For vocabulary learning the implications of these strands of research are that the aid of an understood context can lessen the cognitive load so that higher-level words can be learned through use. Also, a lack of contextual support may make such words unteachable. Without a sufficiently understood context for the new word to cognitively embed itself in, semantic associations would be hard for the learner to create and the "depth of processing" when dealing with the new word would certainly be lower thus decreasing the chances for retention or acquisition.

Such contextual support would hopefully be of particular use to learners with a lower level of vocabulary proficiency because the theme or content would provide hints as to what the unknown language items might be. Below I will give an example of a theme around which a number of tasks could be constructed but first I will describe a framework that will facilitate learner-initiated composition or output using new words.

7.1 The TBL Framework

Laufer and Hulstijn (2001b) mention TBL, in particular the Skehan (1996) model as a sound framework to put the theory and empirical research they aim to stimulate to pedagogical practice. To illustrate how the findings of this research can be utilized, it is suggested to adapt the Willis (1996) framework for TBL to Content-based instruction with an added element of learner-initiated composition.

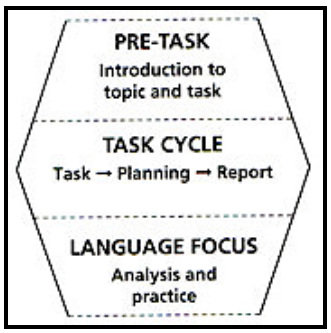

Figure 4. Willis's model of Task-based learning (TBL), Willis (1996, p.60)

As can be seen above, in the Willis framework, after some sort of activity introducing

the topic and task, the students do the task. After that, time is given for the learners to prepare a "report". The effect of planning time on the quality of language production in testing situations has been thoroughly researched. For a summary see Ellis (2003, p. 293). Willis (1996) however suggests the planning time more as a time for the learners to get organized as to what meanings they wish to convey. In a typical group-work or pair-work situation, a report would mean one representative presenting the pair or group's findings to the classroom as a whole. For the current purposes, this report phase can be done twice but in different modes. The first report would be oral and done shortly after task completion. The next report is a written one and shared within groups, then to the class as a whole the next time the group meets. TBL differs from methodological frameworks for language learning such as *"presentation, practice, production*"(PPP) in that, as is the case with most interpretations of communicative language teaching (CLT), for production, the linguistic resources to be used by the learner to complete the task are not to be dictated by the learning materials or the teacher, as Willis (1996, p.24) states "…learners are free to choose whatever language forms they choose to convey what they mean, in order to fulfill, as well as they can, the task goals". Samuda and Bygate (2008) suggest running the task cycle twice in order to introduce the target structure after a meaning space has been created by the first cycle. In the proposed framework, no target structures are introduced. However, as a necessity learners will have to use unfamiliar vocabulary items which they will have encountered earlier in the introduction or task-cycle. These writing assignments can be recycled for a final consciousness-raising activity by lifting common mistakes and focusing on them in accordance with the Willis (1996) precept of focus on form *after* the task cycle. This research has suggested merely using new vocabulary in learner-initiated composition very efficiently promotes acquisition of it. By following a TBL framework with an added written follow-up report phase, aside from the use of language during the task itself, there are two obligatory occasions when learner-initiated composition with new vocabulary items occurs; once during the in-class group report and again for the written report.

7.2 An Example of a Theme

Any number of topics or themes would introduce new vocabulary items particular to the domain, but the following is an example appropriate for "International Education" an

area often dealt with by the EFL Department in high school in Japan. UNESCO's "Decade of Education for Sustainable Development" (DESD) started in 2005 and represents the consensus of what issues the UN member countries think should be addressed in education. In a high school in Japan, courses aimed at such topics would be typically considered "International Education".

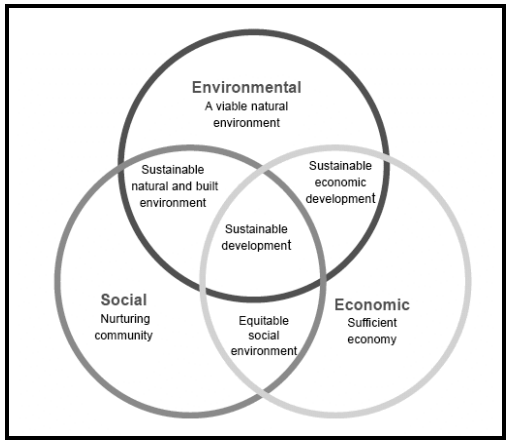

Figure 5. Issues within the realm of "Education for Sustainable Development" (ESD)

As can be seen by the graphical representation, ESD can address economic, environmental, and social issues or those that lie in between. Willis gives examples of 6 types of tasks that can be used in her model of TBL; listing, ordering and sorting, comparing, problem solving, sharing personal experiences, and creative tasks (writing, media projects). It is easy to imagine how to combine TBL with ESD. An example could be as simple as *"Rank the UN's Millennium Development Goals in terms of importance"*. A topic like this will immediately provide a clear meaning space and a plethora of unfamiliar L2 vocabulary items to be used in learner-initiated output in both the oral and written reports. Continuing such a content course with a series of related topics as could be imagined with ESD, the same words would most probably be encountered again giving the learners the chance integrate new words into input as the theme and words associated with it becomes more familiar.

8. Conclusion

As in Laufer and Hulstijn's (2001b) research on their Involvement Load Hypothesis in Amsterdam and Israel with higher level learners, with all other variables equal, for high school learners in Japan, strong evaluation, or learner-initiated composition using new

vocabulary words appears to be more effective at facilitating acquisition than moderate evaluation, or comparing new words with contexts and matching them. An unexpected result however was that the data suggest that the advantages of learner composition may be lower if the learner does not have enough vocabulary knowledge to start out with. The data do not however imply that original composition with new words is less effective than other techniques for lower vocabulary level learners, simply that the advantages are greater if the learners vocabulary level is higher. These results imply that for the language teacher in Japan wishing to aid learners in increasing their L2 lexicon, using a teaching framework that involves learner-initiated composition is effective, advantageous and promotes acquisition. With the goal of encouraging and facilitating learners to use new words in original contexts, one way to do this would be to create courses that adapt TBL to a specific theme or content area so that learners could explore new vocabulary associated particularly with that domain. Providing an overarching theme may aid learners to create their own associations and original contexts by providing a background context to a certain extent. As educators, we owe it to our learners and society to provide the most informed teaching methods we can create.

Appendix 1: Both versions of the worksheets used for the experiment.

Version A: Gap-fill (moderate evaluation)

Child Labor

Is globalization a good thing? Many people think not. Sometimes [1]( ) causes problems. One of them is the [2] ( ) of "child labor".

In some countries, children work and don't go to school. Most work in [3] ( ) for cash [4] ( ). A cash crop is a crop made to sell for money, not to eat or trade with. These crops are usually sold to a foreign country. Sometimes a foreign company owns the [5]( ) the children work at. Some examples of cash crops are coffee, bananas, or [6] ( ). Coconut [7] ( ), or "natadecoco" is a famous cash crop sold to Japan.

Some children work in factories that make things for foreign countries like Japan and the U.S.. Some people call these factories "[8]( ) " (like a hot [9] ( ) because they are so hot and the work is so hard. A [10] ( ) "sweatshop" is a big [11] ( ) factory that makes things for a big company like Nike or Uniclo. Many of the soccer balls we use are made by children. Whether it be cash crops or factories, the children are working to sell things to developed countries. We are [12] ( ) to [13] ( ) for this problem.

Some people in Japan and other countries know about this problem and buy things from fair trade companies like the people tree that don't use child labor and pay the workers better than the "sweatshops". Some people have [14] ( ) against free trade because "free trade" means that more foreign companies can do things like use child labor in developing countries. Because of these demonstrations, some big companies like Nike changed and became fairer, but many haven't yet.

Public opinion can change the world, but first people must be [15] ( ) of the problems. Many students and Free the Children (FTC) will march on Midosuji Sunday 6/8 and will call out to people: "Stop child labor!" At the least, people may notice and think. It is a start. Let's go!

Which words fit in the blanks above?　Write them in the blanks.

crop: 作物	typical: 典型的な	demonstration: デモ	increase: 増加
plantation プランテーション、 大農場	partly： 部分的に、 ある程度	blame: ～の責任にする、 ～のせいにする、	be aware: 気付いて［を知って］ いる
globalization: グローバル化, 経済活 動のグローバル化	agriculture: 農業	clothing: 服、衣類	sweatshop: 労働搾取工場
shrimp:　えび	fiber: 繊維	gymnasium: 体育館	

Comprehension questions: Write a short answer below the questions.

1) What do some people think is not a good thing?

2) Are there more or less children working because of globalization?

3) Where do most of the children work?

4) What is something made at a farm and then sold called?

5) Do the children's families usually own the farms where cash crops are made?

6) What is a seafood that is grown for money?

7) What is a famous cash crop sold to Japan?

8) What is a bad factory called?

9) What is the place behind the image of the word "sweatshop"?

10) Do all "sweatshops" make clothing?

11) Besides shoes, what is Nike known to make?

12) Is child labor 100% the developed countries fault?

13) Is it completely not our fault?

14) What did people do that made Nike change and become better?

15) In order the change these problems, first what must people do?

Name_____ Class/Student Number_____

Version B: Original sentences (strong evaluation)

Child Labor

Is globalization a good thing? Many people think not. Sometimes [1](globalization) causes problems. One of them is the [2](increase) of “child labor”.

In some countries, children work and don't go to school. Most work in [3](agriculture) for cash [4](crops). A cash crop is a crop made to sell for money, not to eat or trade with. These crops are usually sold to a foreign country. Sometimes a foreign company owns the [5](plantation) the children work at. Some examples of cash crops are coffee, bananas, or [6](shrimp). Coconut [7](fiber), or “natadecoco” is a famous cash crop sold to Japan.

Some children work in factories that make things for foreign countries like Japan and the U.S.. Some people call these factories “[8](sweatshops)” (like a hot [9](gymnasium) because they are so hot and the work is so hard. A [10](typical) “sweatshop” is a big [11](clothing) factory that makes things for a big company like Nike or Uniclo. Many of the soccer balls we use are made by children. Whether it be cash crops or factories, the children are working to sell things to developed countries. We are [12](partly) to [13](blame) for this problem.

Some people in Japan and other countries know about this problem and buy things from fair trade companies like the people tree that don't use child labor and pay the workers better than the “sweatshops”. Some people have [14](demonstrations) against free trade because “free trade” means that more foreign companies can do things like use child labor in developing countries. Because of these demonstrations, some big companies like Nike changed and became fairer, but many haven't yet.

Public opinion can change the world, but first people must be [15](aware) of the problems. Many students and Free the Children (FTC) will march on Midosuji Sunday 6/8 and will call out to people: “Stop child labor!“ At the least, people may notice and think. It is a start. Let's go!

Make an original sentence with the following words from the text:

[1] globalization: グローバル化, 経済活動のグローバル化
[2] increase: 増加
[3] agriculture: 農業
[4] crop: 作物
[5] plantation　プランテーション、大農場
[6] shrimp:　えび
[7] fiber: 繊維
[8] sweatshop: 労働搾取工場
[9] gymnasium: 体育館
[10] typical: 典型的な
[11] clothing: 服、衣類
[12] partly：部分的に、ある程度
[13] blame:　～の責任にする、～のせいにする、
[14] demonstration: デモ
[15] be aware: 気付いて［を知って］いる

Comprehension questions: Write a short answer below the questions.

16) What do some people think is not a good thing?

17) Are there more or less children working because of globalization?

18) Where do most of the children work?

19) What is something made at a farm and then sold called?

20) Do the children's families usually own the farms where cash crops are made?

21) What is a seafood that is grown for money?

22) What is a famous cash crop sold to Japan?

23) What is a bad factory called?

24) What is the place behind the image of the word "sweatshop"?

25) Do all "sweatshops" make clothing?

26) Besides shoes, what is Nike known to make?

27) Is child labor 100% the developed countries fault?

28) Is it completely not our fault?

29) What did people do that made Nike change and become better?

30) In order the change these problems, first what must people do?

Name_____ Class/Student Number_____

Appendix 2: Peer selected Unknown Words from the Text

(Numerals indicate number of students who selected the word out of a total of 28 students from the 10th, 11th, and 12th grade.)

Gymnasium 16
Agriculture 16
Cash crops or crops 15
(to be) aware 15
Blame 15
Demonstration 14
Fiber 13
Increase 12
Plantation 10
Warmth 8
Globalization 8
Sweatshop 8
Partly 8
Shrimp 7
Clothing 6
Typical 6
Farming 5
Trade 5
to march 4
Coconut 3
Fair 3
Developed 2
Against 2
Companies 2
Public 2
Whether 1
Developing 1
Causes 1
Nike 1

Whether 1
(child) labor 1
Opinion 1
(at the) least 1

Appendix 3: Pre and Post-tests

How many of these words do you know? Write the translation on the answer sheet.

(pre-version)

1. This is a big gymnasium.
2. Many people work in agriculture.
3. This is where they sells their crops.
4. I am aware of the problem.
5. He is to blame for what happened.
6. There was a demonstration against the new law.
7. This plant has a lot of fiber.
8. Crime has increased in 2007.
9. The coffee is made at a plantation.
10. Some people think globalization is a bad thing.
11. The factory was like a sweatshop.
12. I partly agree with you.
13. His farm makes shrimp.
14. I buy my clothing on the internet.
15. He is a typical university student.

How many of these words do you know? Write the translation on the answer sheet.

(post-version)

1. Many people work in agriculture.
2. He is to blame for what happened.
3. His farm makes shrimp.
4. Some people think globalization is a bad thing.
5. I am aware of the problem.
6. I buy my clothing on the internet.
7. Crime has increased in 2007.
8. I partly agree with you.
9. This is a big gymnasium.
10. This is where they sells their crops.
11. He is a typical university student.
12. The factory was like a sweatshop.
13. This plant has a lot of fiber.
14. There was a demonstration against the new law.
15. The coffee is made at a plantation.

Vocabulary check: Write the Japanese word for the words squared in the sentences.

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	

-pre-

Appendix 4: T-Tests

School A, pre to post-test, control, one-tailed, paired (dependant)	P = 0.08
School A, pre to post-test, gap-fill. one-tailed, paired (dependant)	P = 6.48e-08
School A, pre to post-test, original sentences. one-tailed, paired (dependant)	P = 2.03e-12
School A, post-test to post-test, gap-fill to original sentences, one-tailed, unpaired (independent)	P = 0.01
School B, pre to post-test, control, one-tailed, paired (dependant)	P = 0.37
School B, pre to post-test, gap-fill. one-tailed, paired (dependant)	P = 4.62e-10
School B, pre to post-test, original sentences. one-tailed, paired (dependant)	P = 2.18e-10
School B, post-test to post-test, gap-fill to original sentences, one-tailed, unpaired (independent)	P = 0.39
* number after (e) indicates decimal points to be moved to the left. For example 0.1e-3 would represent 0.0001	

Works Cited:

Altenberg, B, 1990. Speech as linear composition. In: Caie, G. Haastrup, K., Jakobsen, A.L., Neilsen, J.E., Sevaldsen, J., Spect, H. and Zettersten, A. (Eds.) *Proceedings from the Fourth Nordic Conference for English Studies*, Vol. 1, Department of English, University of Copenhagen, 133-143

Atkinson, R.C. (1975) Mnemotechnics in Second-Language Learning. *American Psychologist, Vol. 30:8, 821-828*

Atkinson, R.C., and Shiffrin, R.M. (1968) Human Memory: A proposed system and its control processes. In: K.W. Spence and J.T. Spence (Eds.) The psychology of learning and motivation: *Advances in Research and Theory, Vol. 2, 89-195, New York: Academic Press,*

Baddely, A. (1978) The Trouble with Levels: A Reexamination of Craik and Lockharts Framework for Memory Research. In: *Psychological Review*, *85:3, 139-152*

Breen, M. (1989). The evaluation cycle for language learning tasks. In: R. Johnson (Ed.), *The second language curriculum*. Cambridge, Cambridge University Press.

Brinton, D. (2003). Content-based instruction. In: D. Nunan (Ed.), *Practical English Language Teaching,* 199-224, New York, McGraw Hill.

Carter, R. (1987) *Vocabulary: Applied Linguistic Perspectives*, London, Routledge.

Chomsky, N., (1981). *Lectures on Government and Binding*. Dordrecht, Foris

Chomsky,N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York, Praeger.

Cobb,T. *Web Vocabprofile* [accessed October 2008 from http://www.lextutor.ca/vp/], an adaptation of Heatley & Nation's (1994) Range.

Craik, F.I.M and R.S. Lockhart (1972) 'Levels of Processing: A Framework for Memory Research' In: *Journal of Verbal Learning and Verbal Behaviour* 11:6, 671-84

Cummings, J. (1983) Language proficiency and academic achievement In: J. Oller (ed.). *Issues in Language Testing Research*. Rowely, Mass: Newbury House

Dulay, H.C., & Burt, M.K. (1973). Should we teach children syntax?, *Language Learning*, 23:2, 245-258

Ellis, R.(1994) *The Study of Second Language Acquisition*, Oxford, Oxford University Press.

Ellis, R. (2003) *Task-based Language Learning and Teaching*, Oxford, Oxford University Press.

Ellis, R and He, X (1999) The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, 21, 285-301

Eysenck, M.W. (1982) Incidental Learning and Orienting Tasks, In: C.R. Puff (Ed.): *Handbook of Research Methods in Human Memory and Cognition*. New York: Academic Press. 197-228

Folse, K. S. (2006) The Effect of Type of Written Exercise on L2 Vocabulary Retention, In: *TESOL Quarterly* 40:2, 273-293

Flynn, M. H. (2007) *Electronic Dictionaries, Printed Dictionaries and No Dictionaries: the Effects on Vocabulary Knowledge and Reading Comprehension*, Unpublished Dissertation, University of Birmingam CELS

Gass, S. M. (1988) Integrating research areas: a framework for second language studies, *Applied Linguistics*, 9, 198-217

Gass, S. M. (1999) Discussion: Incidental Vocabulary Acquisition, *Studies in Second Language Acquisition*, 21, 319-333

Hall, E. (1976) *Beyond Culture*, New York, Anchor Books

Heatley, A. and Nation, P. (1994). *Range*. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

Hulstijn, J.H., & Trompetter, P. (1998) Incidental learning of second language vocabulary in computer assisted reading and writing tasks. In: D. Albrechtensen, B. Hendricksen, I.M. Mees, & E. Poulsen (Eds.) *Perspectives on Foreign and Second Language Pedagogy*, 191-200, Odense, Denmark, Odense University Press

Joe, A (1995) Text-based tasks and incidental vocabulary learning: A case study. *Second Language Research*, 11:2, 149-158

Joe, A (1998) What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics*, 19:3, 357-377

Krashen, S. (1982), *Principles and Practice in Second Language Acquisition*, Oxford, Pergamon Press

Laufer, B. (2008a) Conversation with Matthew Walsh, 11th February.

Laufer, B. (2008b) Conversation with Matthew Walsh, 20th February.

Laufer, B. Hulstijn, J. (2001a) Incidental Vocabulary Acquisition in a Second Language: The Construct of Task-Induced Involvement. *Applied Linguistics*, 21:1, 1-26

Laufer, B. Hulstijn, J. (2001b) Some Empirical Evidence for the Involvement Load Hypothesis, *Language Learning*, 51:3, 539-558

Martinez-Fernandez, A. (2008) Revisiting the Involvement Load Hypothesis:

Awareness, Type of Task and Type of Item. In: *Selected Proceedings of the 2007 Second Language Research Forum. (Eds.) Melissa Bowles et al.,* 210-228. Somerville MA, Cascadilla Proceedings Project

Mohan, B. M. (1986). *Language and content*. Reading, MA: Addison-Wesley.

Murdock, B.B. Jr. (1967) Recent Developments in Short-Term Memory. *British Journal of Psychology*, 58:3, 421-433

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary*, Boston: Heinle & Heinle

Nagy, W.E., Herman, P.A. and Anderson, R.C. (1985) 'Learning Words from Context'. *Reading Research Quarterly* 20:2, 233-53.

Paribakt, T & Wesche, M (1997) Vocabulary enhancement activities and reading for meaning in second language acquisition. IN: J. Coady & T. Huckin (Eds.) *Second Language Vocabulary Acquisition*, 174-200, Cambridge, Cambridge University Press

Pinker, S. (1994) *The Language Instinct*, New York: William Morrow and Company.

Pinker, S. (2008) *Charlie Rose,* PBS television March 13th

Richards, J. J. Platt, and H. Weber (1985) *Longman Dictionary of Applied Linguistics*, UK, Longman

Robinson, P. (1995). Attention, memory and the noticing hypothesis. *Language Learning* 45:2, 283-331

Rodriguez, M. and Sadoski, M. (2000) Effects of Rote, Context, Keyword, and Context/Keyword Methods on Retention of Vocabulary in EFL Classrooms. *Language Learning,* 50:2, 385-412

Samuda, V. and Bygate, M. (2008) *Tasks in Second Language Learning*, New York, Palgrave Macmillan

Schmidt (1994) Deconstructing Consciousness in Search of Useful Definitions for Applied Linguistics. IN: J.H. Hulstijn and R. Schmidt (Eds.): *Consciousness in Second Language Learning*. *AILA Review*11, 11-26

Schmidt (2000) Attention In: P. Robinson (Ed.): *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press

Shallice, T., & Warrington, E.K. (1970) Independent Functioning of Verbal Memory Stores: A Neuro-psychological Study. *Quarterly Journal of Experimental Psychology* 22:2, 261-273

Sharwood Smith, M. (1981) Consciousness raising and the second language learner *Applied Linguistics*, 11:2, 159-168

Skehan, P. (1996) A Framework for the Implementation of Task-based Instruction. *Applied Linguistics*, 17:1, 38-62

Skehan, P. (1998) *A Cognitive Approach To Language Learning*. Oxford, Oxford University Press.

Snow, C., M. Met, and F. Genesee. (1989) A conceptual framework for the integration of language and content in second language instruction. *TESOL Quarterly 23:2, 201-217*

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In: S. Gass & C. Madden (Eds.), *Input in second language acquisition*, 235-253, Rowley, MA, Newbury House

Tulving, E, & Patterson, R.D. (1968) Functional Units and Retrieval Processes in Free Recall, In: *Journal of Experimental Psychology*, 77:2, 239-248

Webb, S. (2005) Receptive and Productive Vocabulary Learning: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition*, 27:1, 33-52

Willis, J. (1996) *A Framework For Task-Based Learning*, Essex: Longman

Wray, A., Perkins, M. (2000) "The functions of formulaic language: an integrated model" *Language and Communication, 20:1, 1-28*

Vygotsky, L. (1934) *Thought and Language*, London, MIT Press