Open Journal of Modern Linguistics, 2015, 5, 425-442
Published Online October 2015 in SciRes. http://www.scirp.org/journal/ojml
http://dx.doi.org/10.4236/ojml.2015.55038

Effect of Fine-Grained Lexical Rating on L2 Learners' Lexical Learning Gain

Ruiying Niu

Faculty of English Language and Culture, Guangdong University of Foreign Studies, Guangzhou, China
Email: niuruiying@hotmail.com

Received 25 August 2015; accepted 5 October 2015; published 8 October 2015

Abstract

In assessing L2 lexical learning, especially initial learning, researchers always face the problem of whether partial word learning should be counted. Existing studies have either counted partial word learning (i.e. counted both partial and complete word learning) or have only counted complete word learning. However, it is not clear whether counting partial word learning makes a difference in capturing task-based and intra-learner lexical learning gain. Few studies have investigated this potential difference and even fewer if both productive and receptive lexical learning are considered. The present study employed differently fine-grained word rating methods to assess three Chinese EFL learner groups' performances on four vocabulary posttests after receiving three treatment tasks: a written output task, an oral output task, and a reading task. Data analyses revealed that the use of differently fine-grained scoring methods did not necessarily affect learners' cross-task lexical learning effects significantly, but it did make a significant difference in measuring individual learners' lexical learning gain. The findings are discussed with reference to whether and how a less or more fine-grained scoring method should be adopted in rating lexical learning.

Keywords

L2 Lexical Learning, Multiple Rating Methods, Cross-Task Lexical Learning, Intra-Learner Lexical Learning Gain, Experimental Study

1. Introduction

Second language (L2) vocabulary assessment mainly serves two purposes: one is measuring learners' vocabulary size; the other is gauging learners' vocabulary achievement over a period of time (Bruton, 2009; Read, 2000). For either purpose, learners may manifest partial word learning as lexical learning is incremental (Barcroft & Rott, 2010; Schmitt, 2010). Particularly, in receptive vocabulary tests, learners may provide close

How to cite this paper: Niu, R. Y. (2015). Effect of Fine-Grained Lexical Rating on L2 Learners' Lexical Learning Gain. *Open Journal of Modern Linguistics, 5,* 425-442. http://dx.doi.org/10.4236/ojml.2015.55038

but inaccurate word meanings for target word forms; likewise, in taking a productive vocabulary test, learners may generate inaccurate orthographic (or phonological) word forms. Partial learning is especially common in assessing learners' initial word learning gain in experimental studies (Barcroft & Rott, 2010). An examination of existing studies on lexical learning revealed that some took learners' partial knowledge gain into account (e.g. Barcroft, 2002, 2007; Webb & Chang, 2012), whereas others did not (e.g. Hulstijn & Laufer, 2001; Min, 2008). Bruton (2007) argued that in experimental studies the assessment of incidental vocabulary receptive learning from L2 reading should take into account partial learning since the reading context may not provide learners with adequate lexical learning affordances. Similarly, because of its incremental nature, measures of productive lexical learning should count partial word knowledge as well. The critical point here is that counting partial word knowledge gain might affect cross-task lexical learning impacts and intra-learner lexical knowledge gain while examining the effectiveness of lexical learning tasks. However, few existing studies have specifically addressed this issue. This paper examines whether differently fine-grained lexical learning scoring methods make a difference in determining task-engendered lexical learning as well as individual learners' lexical learning gain through an experimental study.

2. L2 Lexical Knowledge Assessment

Lexical learning assessment pertains to both receptive and productive measures, and these have been defined in different ways. Read (2000) classified receptive and productive vocabulary measures as either recognition (understanding the meaning of an isolated target word) and recall (eliciting the form of a target word based on some stimulus) or comprehension (understanding the meaning of a target word in listening or reading) and use (a target word occurring in speech or writing). As the measures for comprehension and use may be contaminated by contextual factors, the tendency in vocabulary assessment is testing recognition and recall, for example, through L2-to-L1 translations and L1-to-L2 translations respectively (Read, 2000), which actually measure the form-meaning link of an L2 word (Schmitt, 2010). Laufer and Goldstein (2004), based on form-meaning relationships, formulated four degrees of form-meaning knowledge: active recall (supplying L2 word forms), passive recall (supplying L1 equivalents), active recognition (selecting L2 word forms), and passive recognition (selecting L1 equivalents), which Schmitt (2010) named as form recall, meaning recall, form recognition, and meaning recognition for easy understanding. Since form recognition and meaning recognition do not occur in real communications, lexical learning is basically about meaning recall and form recall, the initial steps leading to receptive and productive mastery of words (Schmitt, 2010).

The present study focuses on both receptive and productive lexical learning, which are defined as meaning recall and form recall respectively following the distinctions made in Schmitt (2010). Particularly, receptive lexical learning is operationalized as learners being able to retrieve word meaning when seeing the orthographical form of a target word, and productive lexical learning is operationalized as learners being able to produce the orthographical form of a target word when being provided with the meaning. The methods for assessing receptive and productive lexical learning in previous studies are reviewed below in terms of both testing format and scoring method.

2.1. Assessing Receptive Lexical Learning

Receptive lexical learning defined as meaning recall can be assessed by providing L1 translations, L2 paraphrases, or equivalent pictures for target words depending on the participants' backgrounds. One example is Newton (1995), which required participants with different native language backgrounds to show their understanding of target words in any of the above-mentioned ways. Hulstijn and Laufer (2001) allowed their participants, intact classes of EFL learners, to provide either L1 equivalents or English explanations for target words. Keating (2008), a replication of Hulstijn and Laufer (2001), utilized Spanish-English translation to assess learners' receptive learning of Spanish words.

Such receptive tests focusing on form-meaning links do not incorporate partial word learning in terms of test format, but partial word knowledge can be incorporated in scoring. For example, instead of using a correct/incorrect scoring criterion, Hulstijn and Laufer (2001) and Keating (2008) both considered partially correct meanings by scoring each target word on a three-point scale: 0, 0.5, and 1.

One assessment method emphasizing partial word knowledge in test format is the Vocabulary Knowledge Scale (VKS) (Bruton, 2009; Stewart, Batty, & Bovee, 2012; Wesche & Paribakht, 1996). The VKS is a generic

instrument for measuring the depth or quality of vocabulary knowledge gain. As Wesche and Paribakht (1996: p. 33) claimed, "Its purpose is not to estimate general vocabulary knowledge, but rather to track the early development of specific words in an instructional or experimental situation". The VKS consists of a word knowledge elicitation scale and a scoring scale. The elicitation scale is composed of five categories (Bruton, 2009; Stewart et al., 2012; Wesche & Paribakht, 1996), representing five degrees of word knowledge, going from (I) being completely unfamiliar with a word, (II) having seen a word but not knowing its meaning, (III) being able to guess its meaning (supply synonyms or translations), (IV) knowing its meaning exactly (supply synonyms or translations), to (V) being able to use it (write a sentence). The scoring scale allows for five possible scores: 1, 2, 3, 4 and 5, with 1 and 2 being awarded to test-takers' knowledge degrees I and II respectively, 3 awarded to knowledge degrees III and IV, i.e. when a correct synonym or translation of a word being supplied despite test-takers' certainty, and 4 and 5 awarded to knowledge degree V depending on test-takers using the word in a sentence with semantic appropriateness only or with both semantic appropriateness and grammatical accuracy.

The VKS has been used in a number of studies in its original or modified form. Both Hashemi & Gowdasiaei (2005) and Kim (2008) used the original VKS. The former examined the effects of lexical-set and semantically-unrelated vocabulary instruction on Iranian EFL learners' lexical learning, while the latter, a partial replication of Hulstijn and Laufer (2001), assessed learners' lexical gain from performing reading comprehension, reading plus blank fill-in, and composition writing. Since the VKS has been criticized to be multidimensional, to exclude multiple meanings and approximate word knowledge, and not to really measure productive lexical learning (Bruton, 2009; Stewart et al., 2012), researchers (e.g. Atay & Kurt, 2006; Joe, 1998; Min, 2008; Rott, Williams, & Cameron, 2002; Webb & Chang, 2012) modified the VKS self-report categories for their own purposes. Particularly, Atay & Kurt (2006), instead of using the original five-category VKS report scale, adopted a two-category scale, providing test-takers with an unknown/known option, and required test-takers to supply word meaning and make a sentence with the word if they take the known option. Joe (1998) used the VKS in interviews rather than as a written procedure, to allow for more probing of what the learners know about each word. Joe (1998) also modified the elicitation scale by introducing an extra category, which goes as "I haven't seen this word before, but I think…" so as to allow learners to make inferences about a word through recognizing the prefix, stem, or suffix. Min (2008) condensed the original five VKS categories into four and subsumed them under the basic unknown/known dichotomy, with the unknown dichotomy consisting of Categories I (unknown words) and II (partial word knowledge), and the known dichotomy containing Categories III (supply word meaning) and IV (write a sentence). Rott et al. (2002) only incorporated four VKS categories in order to measure their learners' productive vocabulary learning though the VKS is intended for measuring receptive lexical learning as stated earlier. Webb & Chang (2012) employed three report categories: 0, 1, and 2, respectively for having never seen the word, having seen the word but not knowing its meaning, and supplying word meaning.

Studies may use the VKS test format but not count partial word knowledge at the scoring stage (e.g. de la Fuente, 2002; Min, 2008). De la Fuente (2002) employed an oral receptive version and an oral productive version of the VKS, each containing four self-report categories. However, the author used a correct/incorrect binary scoring procedure, hence losing much of the information that comes from using the VKS. Min (2008) employed four self-report categories but assigned zero scores to Categories I and II, and marked Categories III and IV independently and respectively following a correct/incorrect criterion.

To summarize, the above account of supplying L1 equivalents and the VKS suggests that assessing partial word learning does not so much lie in the test format as in the scoring procedure. The format of supplying L1 equivalents does not entail partial word knowledge, but since learners may not be able to provide exact equivalents, partial word gain can be considered in scoring (Hulstijn & Laufer, 2001; Keating, 2008). The elicitation scale of the VKS provides the possibility for garnering learners' different degrees of word knowledge, which, however, may be ignored in scoring (de la Fuente, 2002; Min, 2008). The point is whether counting partial word knowledge really matters. Particularly, does including partial word knowledge or not really make a difference in assessing cross-task lexical learning effects and intra-learner word knowledge gain? This issue constitutes one focus of this study.

2.2. Assessing Productive Lexical Learning

Productive word learning defined as word form recall can be measured in two test formats. One is isolated

L1-to-L2 word translation (Barcroft & Rott, 2010; Hulstijn & Laufer, 2001) or picture labeling (Barcroft, 2004; de la Fuente, 2002; Ellis & He, 1999; Smith, 2004). The L1 word in L1-to-L2 translation and the picture in picture labeling both serve as meaning prompts for test-takers to recall and produce the target word form so that L1-to-L2 translation and picture labeling can be categorized as the same test format. The L1-to-L2 translation is applicable when test-takers share their L1. However, its problem lies in the possibility of being unable to elicit the right target word. For instance, Barcroft and Rott (2010) conducted immediate posttests in order to elicit target German and Spanish word forms from their participants. The problem with picture labeling is that it is confined to testing some concrete nouns, as seen in previous studies (e.g. Barcroft, 2004; de la Fuente, 2002; Ellis & He, 1999; Smith, 2004).

The other commonly used test format is the productive vocabulary levels test designed by Laufer and Nation (1999). Each test item in this format contains a sentence, which incorporates the target word but leaves its position as a blank. Test-takers are required to fill in the sentential blank so that the target word can be produced. In order to push test-takers to produce the right target word, the initial letter(s) are provided as prompts, as in "The differences were so sl_____ that they went unnoticed" (Schmitt, 2010: p. 203). However, Schmitt (2010) doubts the validity of this test format in that the supplied initial letter(s) and the sentential context might affect the difficulty of the test item. He also argues that such a test format is not communicative enough and cannot measure learners' comprehensive productive word knowledge. Nevertheless, this sentence blank-filling format does involve word form recall, whereas measuring productive vocabulary knowledge via writing or speaking would involve word use instead of word form recall. Furthermore, the blank-filling test tries to overcome the constraints of L1-to-L2 translation and picture labeling as discussed above. However, the use of prompt letters and the sentence context do require validation (Schmitt, 2010). Although this test has been designed to measure productive vocabulary size, the present author considers it equally suitable for measuring learners' productive word gain from task performances, as used in this study.

In spite of their different formats, L1-to-L2 translation (or picture labeling) and the letter-prompted sentence blank-filling both involve word form recall, and may elicit inaccurate word forms, which test formats cannot predict. Yet, inaccurate word forms may be counted in test scoring in order to gain a more accurate picture of learners' lexical gain

However, existing studies focusing on productive lexical learning show discrepancies regarding how partial word knowledge has been counted. Some have followed a correct/incorrect (i.e. 1/0) binary scoring method, which ignores partial word knowledge (de la Fuente, 2002, 2003; Min, 2008). Other studies have adopted a three-point scoring system (i.e. 1, 0.5 and 0) with partial word learning being awarded 0.5 points (e.g. Barcroft, 2009; Barcroft & Sommers, 2005). In addition, Barcroft (2002) has developed a 5-point scoring scale, called the lexical production scoring protocol (LPSP), to measure learners' productive lexical learning as accurately as possible. The LPSP awarded 0.00, 0.25, 0.50, 0.75, or 1.00 to a target word depending on the percentage of "letters correct" (target letters placed in the correct positions of a word) and the percentage of "letters present" (misplaced target letters in a word) in the learner-produced target word. The LPSP has been applied in Barcroft's other studies (Barcroft, 2003, 2007) as well as by other researchers (Keating, 2008; Niu, 2014; Niu & Helms-Park, 2014; Smith, 2004).

Since different scoring approaches count partial lexical knowledge to different degrees, some studies have adopted multiple approaches to capture a fine-grained picture of learners' productive word gain (Barcroft, 2002, 2004). Barcroft (2002), examining the effects of semantic learning, structural learning, versus no elaboration on picture-cued word form recall by L1 English learners of Spanish, applied letter-based scoring, word-based scoring, and syllable-based scoring to measure participants' lexical gain in order to discern the potential scoring moderating effect. Both letter-based scoring and syllable-based scoring counted partial word learning, while word-based scoring did not. However, the study did not observe any significant moderating effect of different scoring approaches on different learning conditions. Barcroft (2004) employed whole word scoring and syllable scoring in measuring the effect of sentence writing and word-meaning repetition on English-speaking Spanish learners' word recall, revealing no significant moderating effect, either. In actuality, neither of the studies intended to particularly examine the different effects of different scoring methods; instead, they validated their research results with different scoring methods. While the two studies revealed that differently fine-grained productive scoring methods did not affect cross-task lexical learning differences, neither examined whether different scoring methods would influence intra-learner lexical learning gain. Therefore, as with receptive lexical learning, more studies are needed to examine whether differently fine-grained productive word scoring methods

make a difference in gauging both cross-task lexical learning effects and individual learners' lexical learning gain.

3. Methodology

3.1. Research Design and Research Questions

Predicated on the above-reviewed literature, the study examined whether counting partial word knowledge would affect cross-task lexical learning differences and intra-learner lexical learning gain. Both receptive and productive lexical learning were included and defined as word meaning recall and orthographic word form recall respectively. The data of the study came from Chinese EFL learners' vocabulary tests taken after performing three input-based tasks: collaborative written output, collaborative oral output, and reading comprehension, sequentially named as Written Output, Oral Output, and Reading in the study. The three tasks were chosen in order to better observe cross-task lexical learning differences because Written Output and Oral Output have been found to be similar but both tend to be significantly better than Reading in bringing about lexical learning (Niu & Helms-Park, 2014).

Counting partial word knowledge was realized via using differently fine-grained scoring methods. Specifically, learners' receptive lexical learning was measured by an adapted VKS, which was scored with two methods: a correct/incorrect binary scoring method called correct meaning scoring and an adapted 5-point scoring method called graded scoring. Graded scoring took learners' partial receptive word knowledge into account while correct meaning scoring did not. It was expected that graded scoring would provide a more accurate picture of learners' receptive lexical learning and hence would reveal cross-task lexical learning effects and individual learners' lexical learning gain more exactly than correct meaning scoring. In order to empirically substantiate such a prediction, two research questions (RQs) were raised in relation to receptive lexical learning:

1) *To what extent do differently fine-grained receptive lexical scoring methods lead to different cross-task lexical learning effects*?

2) *To what extent do differently fine-grained receptive lexical scoring methods lead to different intra-learner lexical learning gain*?

Learners' productive lexical learning was measured by a letter-cued sentence blank-filling following Laufer and Nation (1999), which was scored in three ways: letter-based scoring, syllable-based scoring, and word-based scoring, in light of the methods used in Barcroft (2002, 2004). The three methods counted correct letters, correct syllables, and correct words sequentially. Letter-based scoring is more fined-grained than syllable-based scoring, and syllable-based scoring more fine-grained than word-based scoring. It is predicted that letter-based scoring would reveal a more accurate picture of learners' productive lexical learning and hence would reveal cross-task lexical learning effects and individual learners' lexical learning gain more exactly than syllable-based scoring, and syllable-based scoring would be more exact than word-based scoring in revealing learners' productive lexical learning. While Barcroft (2002, 2004) observed no significant difference among the three scoring methods in revealing cross-task lexical learning effects, no previous studies have investigated whether differently fine-grained scoring methods are related with intra-learner lexical learning gain. Hence, with reference to productive lexical learning, the following two RQs were addressed:

3) *To what extent do differently fine-grained productive lexical scoring methods lead to different cross-task lexical learning effects*?

4) *To what extent do differently fine-grained productive lexical scoring methods lead to different intra-learner lexical learning gain*?

3.2. Participants

240 Chinese year-one English majors from a key university in Guangdong Province, China participated in the study. They were aged 17 - 21, and had studied English for 7 - 11 years. They entered university through passing China's national entrance examinations. In their year-one university study, the participants mainly attended courses relating to the four language skills. They came from 10 intact classes and were divided into three groups: 1) the Written Output group, 2) the Oral Output group, and 3) the Reading group, with 98, 96, and 46 students respectively in each group. The three groups were not significantly different in their overall English proficiency as measured by their term-final English core course scores, $F(2, 237) = 0.206$, $p = 0.814$.

3.3. Reading Input and Target Words

The input passage *The Land of Disney* was selected and adapted from *BBC English*. The passage was expository and hence favorable for learners to remember and recall its content. In order to enhance learners' comprehension, culturally loaded details and non-essential embellished sentences were removed. Although the adapted passage was not difficult for the participants, it contained some relatively difficult words so that the lexical learning purpose could be achieved. The length of the passage (485 words) ensured that the reading task and the post-reading output tasks could be completed within an 80-minute class period. Based on the passage length, the time on task, and the recommended practice in previous studies (e.g. Hulstijn & Laufer, 2001; Laufer, 2003), 10 words plus 6 distracters were selected as target words based on pilot tests on peers of the participants. The target words were all content words, including two nouns (*epoch* and *acuity*), three adjectives (*perilous*, *idyllic* and *apprehensive*) and five verbs (*encapsulate*, *instigate*, *espouse*, *depict* and *heed*). The known word coverage of the passage was approximately 97.94% of the word tokens, which should have ensured the participants' instant comprehension of the passage (Hirsh & Nation, 1992). The target words and distracters were glossed with Chinese meanings on the margin of the passage since word meaning guessing is often unreliable (Laufer, 1997), and learners must first know word meanings if they are to put words to use and store them in memory (e.g. Rott et al., 2002). The marginal Chinese glossing is also coherent with the participants' habit of memorizing English words through associating Chinese equivalents.

3.4. Instruments

3.4.1. Pretest

A word pretest was administered in order to ensure participants' zero baseline knowledge of the target words. The pretest contained the 10 target words and 10 distracters, all key words from the reading input passage. Participants were required to provide Chinese meanings for them. A pretest is believed to be able to gauge learners' word knowledge more accurately than a post-task recall. The pretest effect, if there was any, would apply to all participants, and hence would not affect the research results.

3.4.2. Treatment Tasks

The study employed three treatment tasks, and their detailed requirements are outlined in **Table 1**. The three tasks have been designed to be as comparable as possible in terms of task cycle, time on task, and requirements. Written Output and Oral Output were accompanied by a separate work sheet, on one side of which were the cued words arranged in the order in which they appeared in the passage, and on the other side of which were the cued words ordered alphabetically and followed by their phonetic transcriptions, parts of speech, and Chinese meanings. Cued words were provided in order to remind participants of the propositions of the passage and to optimize their chances of using the target words in reconstruction. The instructions for all three tasks were phrased in both English and Chinese in order to optimize participants' understanding.

3.4.3. Posttests

In light of the productive vocabulary levels test (Laufer & Nation, 1999), a letter-cued sentence blank-filling was adopted to measure participants' productive word learning. It required test-takers to complete a set of blank-embedded sentences, as illustrated below. Both cued letters and Chinese meanings were provided in order

Table 1. Requirements of the three treatment tasks.

Tasks	Stage 1	Stage 2	Stage 3
Written output	Read the input passage individually (10 mins)	Reconstruct the passage in pairs with 16 cued words, and produce a piece of written work (25 mins)	Compare written reconstruction with the original passage in pairs (10 mins)
Oral output		Reconstruct the passage in pairs with 16 cued words orally (15 mins), and then report the reconstruction by one member (10 mins)	Compare oral reconstruction with the original passage in pairs (10 mins)
Reading		Judge whether 16 statements are "true", "false", or "of no evidence" without referring to the passage (25 mins)	Evaluate the correctness of their judgment against the original passage (10 mins)

to prompt learners to generate the target words. The choice of cued letters was determined based on a pilot study. The sentences were chosen and adapted from the British National Corpus in order to optimize their authenticity.

The reality is not as i___c (美好的，愉快的) *as he has expected.*

By drawing on the VKS (Wesche & Paribakht, 1996), an adapted version was utilized to measure participants' receptive lexical gain. It adopted the first four VKS self-report categories, and Categories III and IV required participants to provide Chinese meanings for the tested words instead of providing either translations or synonyms because the input passage was glossed with Chinese. Category V of the original VKS was excluded because it tested word use rather than word recall.

Four posttests, both productive and receptive, were administered in the study. In order to reduce test effects, the four posttests were varied in terms of the number of items and the order in which the items were arranged. Posttest 1 contained 14 items, including 10 target items and 4 distracters, and the 14 items were arranged in the order that they appeared in the input passage. On posttest 2, only the 10 target items remained, and were ordered randomly. Posttests 3 and 4 were the same as posttests 1 and 2 correspondingly. Pilot study analysis indicated that both productive posttests (Alpha = 0.7442; Guttman split-half = 0.6292) and receptive posttests (Alpha = 0.7780; Guttman split-half = 0.8009) were reliable.

3.5. Data Collection Procedure

The data were collected in the following procedure:

1) In Week 1, all three groups completed the vocabulary pre-test within 10 minutes. Then the output group members were paired according to the closeness of their term-final English core course scores and their personal preference. Afterwards, they received a task practice session. The tasks adopted for the practice session were the same as the treatment tasks in terms of both task requirements and task instructions except that a shorter different passage was used and shorter time was allowed for the practice session.

2) In Week 2, all three groups performed their respective treatment task strictly in line with the task instructions. The task performances of the two output groups were audio-taped by the task administrator, who was their listening class teacher, with participants' permission in case that their performances would be useful for interpreting the findings of the study. Recording is a common practice for students in listening classes and should not have affected the participants' performances. Despite being allocated 45 minutes for task completion, Written Output, Oral Output, and Reading averagely took 45, 40 and 30 minutes respectively. This time difference, being regarded as a task-inherent factor, was not considered in data analysis.

3) Immediately upon task completion, all three groups took vocabulary posttests 1, with the productive test preceding the receptive one in order to avoid test effect.

4) In Weeks 3, 5 and 6, vocabulary posttests 2, 3 and 4 were administered.

3.6. Data Analysis

The results of the vocabulary pre-test excluded the target word *apprehensive* from data analysis because some participants happened to be taught this word shortly before the experiment. This word was dropped in order to maintain participants' zero baseline knowledge of target words and maintain an adequate sample size. After the pre-test screening and the exclusion of those absent from any data collection session, 175 participants remained: 69 for Written Output, 72 for Oral Output, and 34 for Reading. One-way ANOVA results of the three groups' term-final English core course scores indicated that they were not significantly different in English proficiency, $F(2, 172) = 1.245$, $p = .29$. Data analysis of the study mainly involved scoring productive and receptive posttests and analyzing the scores statistically in response to the research questions.

3.6.1. Productive Vocabulary Scoring

The three productive vocabulary scoring methods are detailed below.

1) Letter-based scoring

Letter-based scoring calculates the percentage of correct letters in a target word that learners produced and hence takes partially correct word forms into account. Correct word forms are understood as 100% of their letters being correctly produced. The scoring rules were formulated by adapting Barcroft's (2002) LPSP. Applying LPSP to the present study showed that LPSP was not sensitive enough to the variance of incorrect forms. Thus, instead of using a 5-point scale (i.e. 0, 0.25, 0.5, 0.75 and 1), a letter-based scoring system weighed correct let-

ters only, and directly transformed the percentage of correct letters into the score. Three levels were used for letter-based scoring: 1) 0 points were awarded when nothing was written or none of the supplied letters were correct; 2) 1 point was awarded when all supplied letters were correct; 3) a score between 1 and 0 was awarded when a portion of the supplied letters was correct. This was computed through dividing the number of correct letters by the number of target letters or by the number of supplied letters if more letters were written than the target letters. All partially correct words were collected and scored repeatedly. The intrarater scoring reliability based on the author's last two markings was 97.97%. A trained interrater marked all partially correct words and the interrater reliability was 90.36%. Based on letter-based scoring, the score for each target word ranged from 1 to 0. For each participant, the maximum score was 9 and the minimum was 0.

2) Syllable-based scoring

Syllable-based scoring only considers correctly produced syllables. It is less fine-grained than letter-based scoring because scoring partially correct syllables as zero might miss counting some correctly produced letters. Since phonetic syllable divisions and orthographic syllable divisions are not always the same (Wells, 1990), the target words in this study were syllabified on their orthographic forms according to phonetic syllabification principles. The reasons are that the productive posttests required participants to produce orthographic forms of the target words, phonological representations were also involved in participants' task performance, and the orthographic forms of English words are usually connected with their pronunciations (Graddol, 2007; Horobin, 2007). The 9 target words contain totally 24 syllables, 18 of which were incorporated into scoring because the productive posttests were letter-cued and some syllables had been supplied as cued letters. Each correct syllable was awarded 1 point. The total score for all marked syllables ranged from 0 to 18, which were converted into scores ranging from 0 to 9 for statistical analysis in order to be comparable with results from the other two scoring methods.

3) Word-based scoring

Word-based scoring only counts correctly produced words. It is the lease fine-grained method among the three because excluding incorrect word forms means excluding correct letters or correct syllables. Word-based scoring stipulates that one correctly produced word is awarded 1 point; otherwise, 0 points would be assigned. Unanswered items were assigned 0 points because the common practice for productive posttests (sentence blank filling in this study) is that learners leave a blank empty if they do not know the answer. Word-based scoring was conducted by screening the scores resulting from letter-based scoring. In other words, letter-based "1" scores were the "1" scores for word-based scoring, and letter-based "0" scores and partial scores were the "0" scores for word-based scoring.

3.6.2. Receptive Vocabulary Scoring

The two receptive vocabulary scoring methods are stated as follows.

1) Graded scoring

In light of the VKS scoring scale (Wesche & Paribakht, 1996), graded scoring measured participants' receptive word gain on a 5-point scale: 0, 0.25, 0.5, 0.75, and 1. The VKS employed in this study contained four self-report categories. If Category I or Category II was chosen, 0 points or 0.25 points would be awarded respectively. However, three levels of word knowledge were identified for both Category III and Category IV according to the accuracy of the Chinese meaning provided, and each level was assigned a different score. A wrong meaning being supplied equals choosing Category II, and 0.25 points would be awarded. If close meanings were supplied, 0.5 points were awarded. Close meanings refer to meanings partially sharing the original meaning of the target word. For example, the exact Chinese meaning of *acuity* is *minrui* (敏锐), but *jimin* (机敏) would be regarded as the close meaning. To enhance reliability, the close meanings were collected and shown to another Chinese EFL teacher, and agreement was reached between her and the author. If correct meanings were supplied under Category III, the score would be 0.75; if under Category IV, the score would be 1, in order to show participants' degree of certainty in retrieving word meaning. Based on graded scoring, the score for each participant ranged from 9 to 0.

2) Correct meaning scoring

Correct meaning scoring measures receptive vocabulary acquisition solely based on the correctness of the Chinese meanings that participants provided for the target words. This approach is categorical in that either 1 point or 0 points were accorded. One correct Chinese meaning was assigned 1 point; if no meanings, wrong meanings, or close meanings were supplied, 0 points were awarded. In practice, correct meaning scoring was

conducted through counting the items rated at 0.75 points and 1 point in the graded scoring.

3.7. Statistical Analysis

SPSS version 16 was used to analyze the data statistically. In order to answer RQs 1 and 3, one-way ANOVA was applied to compare the differences among the three task groups with reference to letter-based scoring, syllable-based scoring, and whole word scoring respectively for productive lexical learning and with reference to graded scoring and correct meaning scoring respectively for receptive lexical learning. In order to answer RQs 2 and 4, all participants' scores resulting from the three productive word scoring methods were correlated and compared for significant differences with reference to each productive posttest; meanwhile, their scores resulting from the two receptive word scoring methods were correlated and compared for significant differences with reference to each receptive posttest.

4. Results

4.1. Receptive Multiple Scoring and Cross-Task Lexical Learning Effects

As shown in **Table 2**, graded scoring produced higher scores than correct meaning scoring did on all 4 posttests and across all three tasks. This confirms that graded scoring is more fine-grained than correct meaning scoring as assumed in the study.

Table 2. Descriptive statistics from 2 scoring methods on 4 receptive posttests.

Receptive posttest	Scoring methods	Task type	N	Mean	SD	Min.	Max.
1	Graded scoring	Written output	69	6.0525	1.8528	0.75	9.00
		Oral output	72	6.3732	2.1209	0.00	9.00
		Reading	34	5.2190	1.9158	1.25	9.00
	Correct meaning scoring	Written output	69	5.2319	2.2632	0.00	9.00
		Oral output	72	5.6806	2.3606	0.00	9.00
		Reading	34	4.3529	2.2681	0.00	9.00
2	Graded scoring	Written output	69	5.4177	1.9027	1.50	9.00
		Oral output	72	5.9559	1.8735	1.50	9.00
		Reading	34	4.7003	1.9008	0.50	8.25
	Correct meaning scoring	Written output	69	4.5507	2.3673	0.00	9.00
		Oral output	72	5.2361	2.3346	0.00	8.00
		Reading	34	3.5588	2.0478	0.00	9.00
3	Graded scoring	Written output	69	5.5733	2.0799	0.75	9.00
		Oral output	72	6.0556	1.8303	2.00	9.00
		Reading	34	5.1194	2.2073	0.25	9.00
	Correct meaning scoring	Written output	69	4.7246	2.2744	1.00	9.00
		Oral output	72	5.2222	2.2594	1.00	9.00
		Reading	34	4.1471	2.5001	0.00	9.00
4	Graded scoring	Written output	69	5.9525	2.0704	1.25	9.00
		Oral output	72	6.6001	1.9058	1.75	9.00
		Reading	34	5.5125	2.0974	0.00	9.00
	Correct meaning scoring	Written output	69	5.2029	2.5471	0.00	9.00
		Oral output	72	5.8056	2.3355	0.00	9.00
		Reading	34	4.6471	2.5570	0.00	9.00

However, results of one-way ANOVA, as displayed in **Table 3**, revealed scoring-engendered cross-task lexical learning statistical differences only on posttests 1 and 2. That is, Written Output and Reading led to significantly different receptive lexical acquisition on posttest 1 according to graded scoring but not based on correct meaning scoring, whereas the two tasks brought about significantly different receptive lexical retention on posttest 2 according to correct meaning scoring but not based on graded scoring. This finding indicates that differently fine-grained receptive scoring methods, that is, graded scoring and correct meaning scoring in this study, may affect cross-task lexical learning differences significantly. Yet, this effect is variable in that a more fine-grained scoring method is not more likely to produce cross-task differences than a less fine-grained scoring method does, especially when the differences between two tasks are small, like that between Written Output and Reading in this study. However, when two tasks, like Oral Output and Reading used in the study, are sufficiently distant, differently fine-grained scoring methods tend not to significantly affect the degree of their differences in affecting lexical learning. Specifically, as shown in **Table 3**, Oral Output and Reading led to significantly different receptive lexical learning on all 4 posttests based on both scoring methods, hence showing no scoring method effect.

4.2. Receptive Multiple Scoring and Intra-Learner Lexical Learning Gain

As presented in **Table 4**, graded scoring produced higher scores than correct meaning scoring did for all participants (N = 175) on all 4 receptive posttests. This again substantiates that graded scoring can capture fine-grained lexical learning better than correct meaning scoring.

Furthermore, the paired t-test results in **Table 5** reveal that the score differences between graded scoring and correct meaning scoring for all participants reached statistical significance on all 4 posttests. This suggests that the two differently fine-grained receptive scoring methods did produce statistically different intra-learner lexical learning gain. That is, the participants' lexical learning as derived from graded scoring appeared significantly better than that generated from correct meaning scoring. Yet, the score differences resulting from the two

Table 3. Cross-task lexical learning effects of 2 scoring methods on 4 receptive posttests.

Receptive posttest		Task type	Mean difference/MD (Sig.)	
			Written output	Reading
1	Graded scoring	Written Output		0.8335* (0.046)
		Oral Output	0.3207 (0.337)	1.1543** (0.006)
	Correct meaning scoring	Written Output		0.8789 (0.071)
		Oral Output	0.4487 (0.249)	1.3276** (0.002)
2	Graded scoring	Written Output		0.7174 (0.072)
		Oral Output	0.5382 (0.093)	1.2556** (0.002)
	Correct meaning scoring	Written Output		0.9919* (0.041)
		Oral Output	0.6854 (0.078)	1.6773** (0.001)
3	Graded scoring	Written Output		0.4539 (0.282)
		Oral Output	0.4823 (0.156)	0.9362* (0.026)
	Correct meaning scoring	Written Output		0.5776 (0.235)
		Oral Output	0.4976 (0.203)	1.0752* (0.027)
4	Graded scoring	Written Output		0.4399 (0.298)
		Oral Output	0.6476 (0.057)	1.0876* (0.010)
	Correct meaning scoring	Written Output		0.5558 (0.283)
		Oral Output	0.6027 (0.148)	1.1585* (0.025)

Note: *Significant at the $p \leq 0.05$ level; **Significant at the $p \leq 0.01$ level.

Table 4. Descriptive data of all learners resulting from 2 scoring methods on 4 receptive posttests.

Receptive posttest	Scoring method	Mean	N	SD
1	Graded scoring	6.0195	175	2.014
	Correct meaning scoring	5.2794	175	2.3484
2	Graded scoring	5.4834	175	1.9305
	Correct meaning scoring	4.6801	175	2.3547
3	Graded scoring	5.6447	175	1.9968
	Correct meaning scoring	4.9116	175	2.3241
4	Graded scoring	6.0963	175	2.0288
	Correct meaning scoring	5.4482	175	2.4915

Table 5. Mean differences of scores derived from 2 scoring methods on 4 receptive posttests.

Receptive posttest	Scoring method pairs	Mean difference	SD	t	df	Sig.
1	Graded VS correct meaning	0.74004	0.57566	17.006	174	0.000
2	Graded VS correct meaning	0.80323	0.67029	15.852	174	0.000
3	Graded VS correct meaning	0.73315	0.64228	15.1	174	0.000
4	Graded VS correct meaning	0.64804	0.69966	12.253	174	0.000

scoring methods did not change the general trend of participants' lexical learning gain, as shown by the strong correlations between the scores generated from the two scoring methods, as outlined in **Table 6**. That is, those participants who obtained higher scores based on graded scoring tended to gain higher scores, too, according to correct meaning scoring, and *vice versa*.

4.3. Productive Multiple Scoring and Cross-Task Lexical Learning Effects

As displayed in **Table 7**, the scores obtained from letter-based scoring were higher than those derived from syllable-based scoring, and the latter were higher than those generated from word-based scoring for all three tasks and on all four productive posttests. This indicates that letter-based scoring is the most fine-grained scoring method followed by syllable-based scoring while word-based scoring is the least fine-grained among the three, as assumed in this study.

However, results of one-way ANOVA, as shown in **Table 8**, reveal that the three scoring methods did not bring about significantly different cross-task lexical learning on posttests 1 or 3. That is, the lexical learning differences among Written Output, Oral Output, and Reading were consistent in terms of their statistical significance with reference to all three scoring methods. Yet, on posttests 2 and 4, different cross-task lexical learning effects with reference to the three scoring methods were observed. Particularly, on posttest 2 the score differences between Written Output and Reading derived from letter-based scoring and syllable-based scoring were statistically significant, while their score difference generated from word-based scoring did not reach statistical significance. This supports the better fine-grain of letter-based scoring and syllable-based scoring than word-based scoring, which, nevertheless, was not substantiated on posttest 4. Rather, on posttest 4, the score difference between Written Output and Reading reached statistical significance according to syllable-based scoring, but not according to letter-based scoring or word-based scoring. This inconsistency again indicates that differently fine-grained scoring methods may give rise to different cross-task lexical learning effects, but such an effect is variable in that a more fine-grained scoring method does not necessarily more lead to statistically different cross-task lexical learning effects than a less fine-grained scoring method, and *vice versa*.

4.4. Productive Multiple Scoring and Intra-Learner Lexical Learning Gain

As shown in **Table 9**, the participants as a whole (N = 175) obtained the highest average score on all 4

Table 6. Correlations of scores derived from 2 scoring methods on 4 receptive posttests.

Receptive posttest	Scoring methods	Correct meaning scoring
1	Graded scoring	0.977 (N = 175, p = 0.000)
2	Graded scoring	0.970 (N = 175, p = 0.000)
3	Graded scoring	0.967 (N = 175, p = 0.000)
4	Graded scoring	0.973 (N = 175, p = 0.000)

Table 7. Descriptive statistics from 3 scoring methods on 4 productive posttests.

Productive posttest	Scoring method	Task type	N	Mean	SD	Min.	Max.
1	Letter-based scoring	Written output	69	4.7926	2.3043	0.67	9.00
		Oral output	72	5.4356	2.3366	0.00	9.00
		Reading	34	2.9812	2.0489	0.00	7.21
	Syllable-based scoring	Written output	69	3.8406	2.5560	0.00	9.00
		Oral output	72	4.5208	2.5980	0.00	9.00
		Reading	34	2.3235	1.9882	0.00	7.00
	Word-based scoring	Written output	69	3.4058	2.5914	0.00	9.00
		Oral output	72	3.9444	2.4888	0.00	9.00
		Reading	34	2.0000	1.7922	0.00	6.00
2	Letter-based scoring	Written output	69	2.3859	1.6207	0.00	7.61
		Oral output	72	2.7264	1.7821	0.00	8.55
		Reading	34	1.5429	1.3318	0.00	5.24
	Syllable-based scoring	Written output	69	1.6594	1.5351	0.00	6.50
		Oral output	72	1.9097	1.7609	0.00	8.00
		Reading	34	0.9559	1.0028	0.00	4.00
	Word-based scoring	Written output	69	1.2754	1.4741	0.00	6.00
		Oral output	72	1.5694	1.6515	0.00	7.00
		Reading	34	0.6765	1.0363	0.00	4.00
3	Letter-based scoring	Written output	69	2.5393	2.0698	0.00	8.00
		Oral output	72	2.9543	1.8361	0.00	9.00
		Reading	34	1.8076	1.4027	0.00	4.97
	Syllable-based scoring	Written output	69	1.8406	1.9355	0.00	7.00
		Oral output	72	1.9375	1.7681	0.00	9.00
		Reading	34	1.1471	1.0768	0.00	4.00
	Word-based scoring	Written output	69	1.6087	1.9267	0.00	7.00
		Oral output	72	1.7778	1.7541	0.00	9.00
		Reading	34	0.9412	0.8856	0.00	3.00
4	Letter-based scoring	Written output	69	3.4968	2.4066	0.00	9.00
		Oral output	72	3.7761	2.2848	0.00	9.00
		Reading	34	2.5994	2.1758	0.00	9.00
	Syllable-based scoring	Written output	69	2.7609	2.4624	0.00	9.00
		Oral output	72	2.7708	2.2344	0.00	9.00
		Reading	34	1.7941	2.0192	0.00	9.00
	Word-based scoring	Written output	69	2.5507	2.5178	0.00	9.00
		Oral output	72	2.6528	2.2961	0.00	9.00
		Reading	34	1.7647	2.0160	0.00	9.00

Table 8. Cross-task lexical learning effects of three scoring methods on 4 productive posttests.

Productive posttest	Scoring method	Task type	Mean difference/MD (Sig.)	
			Written output	Reading
1	Letter-based scoring	Written output		1.8114** (0.000)
		Oral output	0.6430 (0.095)	2.4544** (0.000)
	Syllable-based scoring	Written output		1.5171** (0.004)
		Oral output	0.6803 (0.105)	2.1973** (0.000)
	Word-based scoring	Written output		1.4058** (0.006)
		Oral output	0.5387 (0.187)	1.9444** (0.000)
2	Letter-based scoring	Written output		0.843* (0.015)
		Oral output	0.3405 (0.220)	1.1835** (0.001)
	Syllable-based scoring	Written output		0.7035* (0.032)
		Oral output	0.2503 (0.339)	0.9538** (0.004)
	Word-based scoring	Written output		0.5989 (0.055)
		Oral output	0.2941 (0.240)	0.8930** (0.004)
3	Letter-based scoring	Written output		0.7316 (0.062)
		Oral output	0.4150 (0.187)	1.1467** (0.004)
	Syllable-based scoring	Written output		0.6935 (0.057)
		Oral output	0.0969 (0.740)	0.7904* (0.029)
	Word-based scoring	Written output		0.6675 (0.063)
		Oral output	0.1691 (0.556)	0.8366* (0.019)
4	Letter-based scoring	Written output		0.8974 (0.066)
		Oral output	0.2793 (0.475)	1.1767* (0.016)
	Syllable-based scoring	Written output		0.9668* (0.045)
		Oral output	0.0100 (0.979)	0.9767* (0.042)
	Word-based scoring	Written output		0.7860 (0.110)
		Oral output	0.1021 (0.796)	0.8881 (0.070)

Note: *Significant at the $p \leq 0.05$ level; **Significant at the $p \leq 0.01$ level.

Table 9. Descriptive data of all learners resulting from 3 scoring methods on 4 productive posttests.

Productive posttest	Scoring method	Mean	N	SD
1	Letter-based scoring	4.7052	175	2.42973
	Syllable-based scoring	3.8257	175	2.58803
	Word-based scoring	3.3543	175	2.50291
2	Letter-based scoring	2.3622	175	1.68674
	Syllable-based scoring	1.6257	175	1.58066
	Word-based scoring	1.28	175	1.50722
3	Letter-based scoring	2.5679	175	1.8975
	Syllable-based scoring	1.7457	175	1.74599
	Word-based scoring	1.5486	175	1.71762
4	Letter-based scoring	3.4374	175	2.3404
	Syllable-based scoring	2.5771	175	2.3084
	Word-based scoring	2.44	175	2.3478

productive posttests based on letter-based scoring, the lowest average score based on word-based scoring, and the medium average score according to syllable-based scoring. This also substantiates the assumption of the study that letter-based scoring is the most fine-grained method, word-based scoring the least fine-grained, and syllable-based scoring in the middle.

Results of paired-samples t-test, as displayed in **Table 10**, reveal that the score differences generated from the three scoring methods for all participants were all statistically significant on all 4 productive posttest. This suggests that differently fine-grained scoring methods were able to lead to learners' significantly different lexical learning gain. That is, the participants' productive lexical learning as measured by letter-based scoring appeared significantly better than that produced by syllable-based scoring, and their lexical learning as measured by syllable-based scoring was significantly better than that generated by word-based scoring. However, the score differences resulting from different scoring methods did not change the general trend of the participants' lexical learning gain as shown by the correlations in **Table 11**. That is, the participants who garnered higher scores based on letter-based scoring tended to obtain higher scores according to syllable-based scoring and word-based scoring, too, and *vice versa*.

Table 10. Mean differences of scores derived from 3 scoring methods on 4 productive posttests.

Productive posttest	Scoring method pairs	Mean	SD	t	df	Sig.
1	Letter-based VS word-based	1.3509	1.0511	17.003	174	0.000
	Syllable-based VS word-based	0.4714	0.8869	7.032	174	0.000
	Letter-based VS syllable-based	0.8795	0.76604	15.188	174	0.000
2	Letter-based VS word-based	1.0822	0.74155	19.036	174	0.000
	Syllable-based VS word-based	0.3457	0.6492	7.045	174	0.000
	Letter-based VS syllable-based	0.7365	0.60185	16.189	174	0.000
3	Letter-based VS word-based	1.0193	0.8221	16.402	174	0.000
	Syllable-based VS word-based	0.1971	0.66501	3.922	174	0.000
	Letter-based VS syllable-based	0.8222	0.63429	17.147	174	0.000
4	Letter-based VS word-based	0.9974	0.76861	17.166	174	0.000
	Syllable-based VS word-based	0.1371	0.6442	2.816	174	0.005
	Letter-based VS syllable-based	0.8602	0.67088	16.962	174	0.000

Table 11. Correlations of scores derived from 3 scoring methods on 4 productive posttests.

Productive posttest	Scoring methods	Syllable-based scoring	Word-based scoring
1	Letter-based scoring	0.955 (0.000)	0.910 (0.000)
	Syllable-based scoring		0.940 (0.000)
2	Letter-based scoring	0.934 (0.000)	0.898 (0.000)
	Syllable-based scoring		0.913 (0.000)
3	Letter-based scoring	0.943 (0.000)	0.901 (0.000)
	Syllable-based scoring		0.926 (0.000)
4	Letter-based scoring	0.958 (0.000)	0.946 (0.000)
	Syllable-based scoring		0.962 (0.000)

5. Discussion

The study found that for both receptive and productive lexical learning, differently fine-grained scoring methods might not necessarily lead to significantly different cross-task effects (in response to RQs 1 and 3), whereas they did bring about significantly different intra-learner word gain (in response to RQs 2 and 4). The finding about the association between differently fine-grained scoring methods and cross-task productive lexical learning effects substantiates that of Barcroft (2002, 2004). As no previous studies had investigated whether and how differently fine-grained scoring methods would affect cross-task receptive lexical learning effects or intra-learner lexical learning gain, the present study bridges the gaps and makes contributions in these respects. The findings are discussed in terms of the association between multiple rating methods and cross-task effects as well as intra-learner lexical learning gain respectively.

5.1. Multiple Ratings and Cross-Task Lexical Learning Effects

The finding that differently fine-grained scoring methods did not consistently lead to significantly different cross-task receptive or productive lexical learning effects indicates that the score gap resulting from different scoring methods might not be large enough to make significant differences. This should be attributed to the fact that a scoring method, if applied, had been applied consistently to rate the lexical learning of all three task groups. Regarding productive lexical learning, because of their increasing fine-grain, word-based scoring, syllable-based scoring, and letter-based scoring produced increasingly higher scores for all three groups. In other words, the Oral Output group obtained higher scores according to letter-based scoring than they did according to syllable-based scoring or word-based scoring, which was also true with the Written Output group and the Reading group. In this case, it is natural that the cross-task lexical learning effects derived from the three scoring methods might not be significantly different, especially when the lexical learning effects of different tasks were adequately close. Similarly, in measuring receptive lexical learning, graded scoring produced higher scores for all three groups than correct meaning scoring did since the former is more fine-grained than the latter. Thus, it is also natural that the cross-task receptive lexical learning effects generated from the two scoring methods might be so close as to have no significant differences.

The finding about the association between multiple ratings and cross-task effects implies that in task-based experimental studies, it may not matter much whether a more or less fine-grained scoring method is used in rating lexical learning as long as a scoring method is used consistently. With regard to productive lexical learning, one evidence comes from Barcroft (2002, 2004), which applied differently fine-grained lexical learning scoring methods but did not find significant cross-task moderating effects. The finding of the present study further confirms that of Barcroft (2002, 2004). The literature provides various productive lexical rating options such as binary scoring (e.g. de la Fuente, 2003), 3-point scoring (e.g. Barcroft, 2009; Barcroft & Sommers, 2005), or 5-point scoring (e.g. Barcroft, 2003; Barcroft, 2007; Keating, 2008; Smith, 2004). The first one did not consider partial word learning while the latter two did. The finding of the study indicates that researchers can choose any of the above methods in examining task-based lexical learning differences, since these discrepant scoring methods do not usually significantly affect cross-task effects. Likewise, with respect to receptive lexical learning, it is also not essential whether partial word meaning recall is considered (e.g. Hulstijn & Laufer, 2001; Keating, 2008) or not (e.g. de la Fuente, 2002; Min, 2008). The VKS aims to count learners' partial word learning (Wesche & Paribakht, 1996). The finding of the present study indicates that it is worth examining the necessity of using the VKS in measuring task-related receptive lexical learning.

In actuality, the use of a more or less fine-grained scoring method in rating lexical learning depends on how word learning is defined because differently fine-grained scoring methods target different components that learners recall. In the present study, letter-based scoring, syllable-based scoring, and word-based scoring sequentially take the correct letter, the correct syllable, and the correct word as the scoring unit. A score of 5 based on letter-based scoring or syllable-based scoring does not necessarily mean that the learner has recalled 5 target words, while a score of 5 based on word-based scoring means that the learner has recalled 5 words correctly. Thus, letter-based scoring and syllable-based scoring might be misleading if productive word learning is defined as correctly producing a word. In like manner, a score of 1 generated from graded scoring does not necessarily mean that the learner recalled the meaning of a word correctly as correct meaning scoring did. Instead, it might be that the learner selected Category II of the VKS for four different words. Hence, graded scoring is misleading, too, if receptive lexical learning is defined as recalling the meaning of a word correctly. Therefore, researchers

need to consider the conceptualizations behind differently fine-grained scoring methods when considering rating partial word learning.

5.2. Multiple Ratings and Intra-Learner Lexical Learning Gain

The study found that for both receptive and productive lexical learning, differently fine-grained scoring methods brought about significantly different intra-learner word learning gain. This finding implies that whether partial word learning is counted or not will matter a lot in judging the effectiveness of a task through measuring learners' lexical learning, which can be extended to measuring learners' vocabulary size through administering vocabulary tests. Specifically, it is very likely that counting partial word learning or not (for instance, employing letter-based or syllable-based scoring rather than word-based scoring in rating productive lexical learning, and adopting graded scoring instead of correct meaning scoring in rating receptive lexical learning) will reveal significantly different degrees of task-related lexical gain (and significantly discrepant vocabulary sizes for individual learners). Therefore, researchers should be very cautious in choosing rating methods for measuring individual learners' lexical gain (or vocabulary size). Particularly, if they aim to capture learners' lexical gain over a period of time (or vocabulary size) as accurately as possible, such partial-learning-incorporated scoring systems as the VKS scoring scheme (Wesche & Paribakht, 1996) and the LPSP (Barcroft, 2002) are indispensible.

However, the problem of defining lexical learning also applies here since the use of differently fine-grained scoring methods may mean different lexical units being measured. Considering the connotations behind the differently fine-grained scoring methods as discussed above, researchers should be clear about the essential differences that using different rating methods may make in measuring individual learners' lexical learning gain (or vocabulary size). On the one hand, choosing differently fine-grained scoring methods means that different degrees of lexical gain (or different vocabulary sizes) may be garnered for individual learners; on the other hand, differently fine-grained scoring methods may mean different word components being weighed, as argued above. Thus, researchers should carefully ensure that their definition of lexical learning and what they measure are consistent.

6. Conclusion

The study interestingly revealed that the use of differently fine-grained scoring methods might not necessarily affect cross-task lexical learning effects significantly, but it did make a difference in measuring individual learners' lexical learning. The study not only bridges a research gap about the association between differently fine-grained scoring methods and learners' lexical gain, but its findings provide both empirical evidence and implications for whether and how a more or less fine-grained scoring method should be adopted in rating lexical learning. Like many studies, the study also has its limitations. Specifically, it only used a small word sample, and did not control the length of the words in terms of the number of letters or syllables that they contained, which might affect the results of the study. Hence, future research may examine a larger word sample and use words of the same length to validate the findings of the study.

Acknowledgements

This study has been supported by China's Educational Ministry humanity social science key research center project (No. 12JJD740006); and also supported by Guangdong Province social science 12th 5-year discipline construction project (No. GD12XWW02).

References

Atay, D., & Kurt, G. (2006). Elementary School EFL Learners' Vocabulary Learning: The Effects of Post-Reading Activities. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes, 63*, 255-273. http://dx.doi.org/10.3138/cmlr.63.2.255

Barcroft, J. (2002). Semantic and Structural Elaboration in L2 Lexical Acquisition. *Language Learning, 52*, 323-363. http://dx.doi.org/10.1111/0023-8333.00186

Barcroft, J. (2003). Effects of Questions about Word Meaning during L2 Spanish Lexical Learning. *Modern Language Journal, 87*, 546-561. http://dx.doi.org/10.1111/1540-4781.00207

Barcroft, J. (2004). Effects of Sentence Writing in Second Language Lexical Acquisition. *Second Language Research, 20*,

303-334. http://dx.doi.org/10.1191/0267658304sr233oa

Barcroft, J. (2007). Effects of Opportunities for Word Retrieval during Second Language Vocabulary Learning. *Language Learning, 57,* 35-56. http://dx.doi.org/10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2009). Effects of Synonym Generation on Incidental and Intentional L2 Vocabulary Learning during Reading. *TESOL Quarterly, 43,* 79-103.

Barcroft, J., & Rott, S. (2010). Partial Word Form Learning in the Written Mode in L2 German and Spanish. *Applied Linguistics, 31,* 623-650. http://dx.doi.org/10.1093/applin/amq017

Barcroft, J., & Sommers, M. (2005). Effects of Acoustic Variability on Second Language Vocabulary Learning. *Studies in Second Language Acquisition, 27,* 387-414. http://dx.doi.org/10.1017/s0272263105050175

Bruton, A. (2007). Partial Lexical Learning in Tests of Incidental Vocabulary Learning from L2 Reading. *The Canadian Modern Language Review, 64,* 163-180. http://dx.doi.org/10.3138/cmlr.64.1.163

Bruton, A. (2009). The Vocabulary Knowledge Scale: A Critical Analysis. *Language Assessment Quarterly, 6,* 288-297. http://dx.doi.org/10.1080/15434300902801909

de la Fuente, M. J. (2002). Negotiation and Oral Acquisition of L2 Vocabulary: The Roles of Input and Output in the Receptive and Productive Acquisition of Words. *Studies in Second Language Acquisition, 24,* 81-112. http://dx.doi.org/10.1017/S0272263102001043

de la Fuente, M. J. (2003). Is SLA Interactionist Theory Relevant to CALL? A Study on the Effects of Computer-Mediated Interaction in L2 Vocabulary Acquisition. *Computer-Assisted Language Learning, 16,* 47-81. http://dx.doi.org/10.1076/call.16.1.47.15526

Ellis, R., & 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. http://dx.doi.org/10.1017/s0272263199002077

Graddol, D. (2007). English Manuscripts: The Emergence of a Visual Identity. In S. Goodman, D. Graddol, & T. Lillis (Eds.), *Redesigning English* (pp. 161-204). London: The Open University.

Hashemi, M. R., & Gowdasiaei, F. (2005). An Attribute-Treatment Interaction Study: Lexical-Set versus Semantically-Unrelated Vocabulary Instruction. *RELC Journal, 36,* 341-361. http://dx.doi.org/10.1177/0033688205060054

Hirsh, D., & Nation, P. (1992). What Vocabulary Size Is Needed to Read Unsimplified Texts for Pleasure? *Reading in a Foreign Language, 8,* 689-696.

Horobin, S. (2007). *Chaucer's English*. New York: Palgrave MacMillan.

Hulstijn, J., & Laufer, B. (2001). Some Empirical Evidence for the Involvement Load Hypothesis in Vocabulary Acquisition. *Language Learning, 51,* 539-558. http://dx.doi.org/10.1111/0023-8333.00164

Joe, A. (1998). What Effects Do Text-Based Tasks Promoting Generation Have on Incidental Vocabulary Acquisition? *Applied Linguistics, 19,* 357-377. http://dx.doi.org/10.1093/applin/19.3.357

Keating, G. D. (2008). Task Effectiveness and Word Learning in a Second Language: The Involvement Load Hypothesis on Trial. *Language Teaching Research, 12,* 365-386. http://dx.doi.org/10.1177/1362168808089922

Kim, Y. (2008). The Role of Task-Induced Involvement and Learner Proficiency in L2 Vocabulary Acquisition. *Language Learning, 58,* 285-325. http://dx.doi.org/10.1111/j.1467-9922.2008.00442.x

Laufer, B. (1997). The Lexical Plight in Second Language Reading: Words You Don't Know, Words You Think You Know and Words You Can't Guess. In J. Coady, & T. Huckin (Eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 20-34). Cambridge: Cambridge University Press.

Laufer, B. (2003). Vocabulary Acquisition in a Second Language: Do Learners Really Acquire Most Vocabulary by Reading? Some Empirical Evidence. *The Canadian Modern Language Review, 59,* 567-587. http://dx.doi.org/10.3138/cmlr.59.4.567

laufer, B., & Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning, 54,* 399-436. http://dx.doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Nation, P. (1999). A Vocabulary-Size Test of Controlled Productive Ability. *Language Testing, 16,* 33-51. http://dx.doi.org/10.1177/026553229901600103

Min, H.-T. (2008). EFL Vocabulary Acquisition and Retention: Reading plus Vocabulary Enhancement Activities and Narrow Reading. *Language Learning, 58,* 73-115. http://dx.doi.org/10.1111/j.1467-9922.2007.00435.x

Newton, J. (1995). Task-Based Interaction and Incidental Vocabulary Learning: A Case Study. *Second Language Research, 11,* 159-177. http://dx.doi.org/10.1177/026765839501100207

Niu, R. (2014). Chinese EFL Learners' Actual Word Processing and Lexical Learning in Performing a Collaborative Output Task. *Chinese Journal of Applied Linguistics, 37,* 309-333. http://dx.doi.org/10.1515/cjal-2014-0020

Niu, R., & Helms-Park, R. (2014). Interaction, Modality, and Word Engagement as Factors in Lexical Learning in a Chinese Context. *Language Teaching Research, 18,* 345-372. http://dx.doi.org/10.1177/1362168813510383

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511732942

Rott, S., Williams, J., & Cameron, R. (2002). The Effect of Multiple-Choice L1 Glosses and Input-Output Cycles on Lexical Acquisition and Retention. *Language Teaching Research, 6*, 183-222. http://dx.doi.org/10.1191/1362168802lr108oa

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. New York: Palgrave Macmillan. http://dx.doi.org/10.1057/9780230293977

Smith, B. (2004). Computer-Mediated Negotiated Interaction and Lexical Acquisition. *Studies in Second Language Acquisition, 26*, 365-398. http://dx.doi.org/10.1017/s027226310426301x

Stewart, J., Batty, A. O., & Bovee, N. (2012). Comparing Multidimensional and Continuum Models of Vocabulary Acquisition: An Empirical Examination of the Vocabulary Knowledge Scale. *TESOL Quarterly, 65*, 695-721.

Webb, S., & Chang, A. C.-S. (2012). Vocabulary Learning through Assisted and Unassisted Reading. *Canadian Modern Language Review, 68*, 267-290. http://dx.doi.org/10.3138/cmlr.1204.1

Wells, J. C. (1990). *Longman Pronunciation Dictionary*. London: Longman Group UK Limited.

Wesche, M., & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge: Depth vs. Breadth. *Canadian Modern Language Review, 53*, 13-39.