Journal of Humanities and Social Sciences, Vol. 6, No. 1, pp. 87-108 (2009)

When One is Not Enough: Translation Rating and the Assessment of Partial Word Knowledge

CAROL TROY and SYOU-RUNG TSAU
Department of English Language, Da-Yeh University
No. 168, University Rd., Dacun, Changhua 51591, Taiwan, R.O.C.

ABSTRACT

An important requirement in foreign language incidental vocabulary acquisition research is accurate vocabulary assessment of partial word knowledge. Open-format L1 translation tests are increasingly used for this purpose. What level of precision is appropriate in the translation rating procedure? To answer this question, experimental data from a read-and-test study are analyzed. The pretest and posttest translations are rated on an eleven-level scale. Through an approximation process, equivalent binary, three-level, and six-level data are derived. The Mann-Whitney U Test is applied to each of the four data sets (eleven-level, binary, three-level, and six-level) to identify the words for which subject knowledge improvement reached significance. By using the original, eleven-level data as a standard, it is shown that binary and three-level rating lead to false positives and false negatives. Two conclusions are drawn: 1. Not all partially correct translations deserve equal credit; and 2. Multi-level rating is a more precise measure of translation accuracy than binary and three-level rating. Practical rating issues and the advantages of using a pretest and posttest as opposed to a posttest only are also discussed.

***Key Words*:** incidental vocabulary acquisition, partial word knowledge, translation rating, vocabulary assessment

當一種方法不足時：字彙譯解及部分知識之評量

杜凱蕾　曹秀蓉
大葉大學英美語文學系
51591 彰化縣大村鄉學府路 168 號

摘　要

在非刻意字彙習得研究裡，對於部分字彙知識作正確的評量是重要且必需的，開放式的母語譯解愈來愈常被用來作此類的評量，何種精確度適用於字彙譯解評量？爲了回答這個問題，我們分析一份「字彙譯解前測→閱讀→字彙譯解後測」的實驗數據。我們使用 11 點量表評量字彙譯解前測與後測。接著經由一個概算程序，我們得到對應的 2 點、3 點及 6 點量表資料。我們採用曼惠特寧 U 法檢定這四組資料（即 11 點、2 點、3 點及 6 點）以便找出受測者在前後測的字彙知識進步達到顯著水準的那些單字。以最原始的 11 點量表爲標準，我們發現 2 點及 3 點量表造成錯誤的正數及負數。依此我們得到兩個結論：1. 並非所有部分正確的字彙譯解可以

得到相同的分數；2. 多點量表比 2 點及 3 點量表能更準確評出字彙譯解的成績。我們進而探討實用的評量方法及使用前後測比只使用後測有更多的優點。

關鍵詞：非刻意字彙習得，專業英語，部分字彙知識，譯解評量，字彙評量

I. INTRODUCTION

Learners acquire vocabulary incidentally when they pick up new words by reading or listening without "trying" to do so. It is well known that incidental reading exposure leads to impressive L1 vocabulary growth among school children (Nagy, Herman, & Anderson, 1985). Several studies show that foreign language (FL) learners also pick up new word meanings incidentally through reading (Day, Omura, & Hiramatsu, 1991; Rott, 1999; Webb, 2007). However, none of them has shown that L2 incidental acquisition through reading produces the same durable long term vocabulary growth experienced by L1 learners. Clearly, L1 and FL vocabulary growth follow different models.

Current research is giving us a better understanding of L2 vocabulary acquisition by exploring the underlying variables, such as the number of reading exposures (Rott, 1999; Webb, 2007), the features of the word contexts (Webb, 2007), and enhancement through teacher support (Ulanoff & Pucci, 1999). All of this research requires the gathering and analysis of experimental data. Horst (2000) has coined the term "read-and-test" to describe the usual format of such studies: She describes a typical read-and-test study as follows:

> Participants read a text that contains words researchers have targeted for learning, but the participants do not know this. They read the text in the normal way, that is, they read to comprehend its informational content. After they finish reading, the participants take an unexpected test of knowledge of the target words. (Horst, 2000, Chapter 3)

The slowness of FL incidental acquisition (Hunt & Beglar, 2005) poses a number of methodological problems for the design of such posttests. One is accurate assessment of partial word knowledge (PWK). Acquisition of new word meanings is known to occur in increments, over a period of time. In FL acquisition, the length of this period may exceed the time frame of the study. As shown in Figure 1 below, the learner may not be 100% unfamiliar with the meaning of a target word at the time of the pretest, and (s)he may not achieve full knowledge of its meaning (FWK, or full word knowledge) by the time of the posttest.[1] As shown in the gray area, the

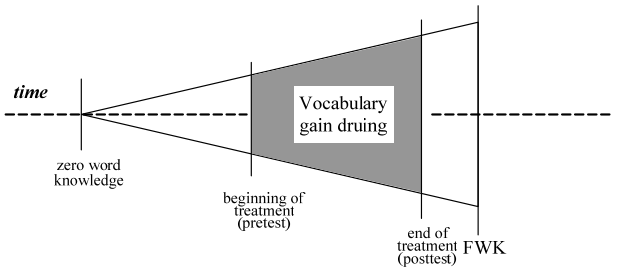


Fig. 1. Growth in word knowledge vs. experimental time frame

resulting vocabulary gain is incomplete, or partial. Current methods of vocabulary assessment are not sensitive to PWK (Horst, 2000), and therefore in practice it is difficult to detect such partial gains in vocabulary tests.

Can open format translation tests, in which the subject translates a list of L2 target words into L1 and each translation is rated for accuracy, detect incremental improvements in word knowledge? On one hand, open format tests have the following three advantages over other test formats in the measurement of PWK.

First, translations are limited in scope. They reflect a single aspect of word knowledge, namely passive knowledge of word *meanings*. Recent innovations in vocabulary testing focus on the diverse nature of word knowledge. Full word knowledge includes not only knowledge of word meanings, but also of orthography, appropriate productive use the word, and other word features. Seen from this perspective, any type of missing knowledge (e.g., ignorance of pronunciation) constitutes PWK. Translation tests are interesting because they give us an alternate view of PWK, in which we may observe incremental accumulation in a single area of word knowledge. As a reflection of knowledge of word meaning, translations are an appropriate choice. It is known that they are employed in the development of L2 lexical entries, and learners – particularly those limited proficiency – rely on them for access to L2 word meanings. Moreover, translation tests, unlike multiple choice and cloze tests, do not require the use of L2 reading passages, which may influence the subjects' performance in ways that are not strictly vocabulary related.

Second, translation tests provide flexibility in scoring.

[1] This diagram is intended as an abstract, schematic depiction of word growth. In practice, changes in word knowledge may not follow an increasing linear path. For example, in the case of a misleading word context, knowledge may decline.

They do not require *a priori* design decisions that constrain the scoring strategy. Writing a translation test involves only selecting a list of appropriate L2 target words. Though the general approach to scoring may be chosen in advance, the implementation details can be deferred until after pretest and posttest data collection, when the investigator can make adjustments based on the subjects' responses. For example, suppose that the subjects misinterpret a particular target word in some unanticipated way, leading to partially correct translation responses. Then the translation rating criteria for that word could be written to identify the subjects who exhibit this gap in understanding. It would be difficult or impossible to make such adjustments in a closed form test (e.g., a multiple choice or cloze test) because the investigator has already decided up front, before data collection, what detailed information to gather. These design choices limit the investigators' flexibility: once the target data have been specified, there is little room for change. By contrast, translation responses contain rich information reflecting the subjects' understanding of the word meaning, which remains available to the investigator during scoring. After testing, the rating criteria can be calibrated to extract the most pertinent information for analysis. We will see examples of this versatility below.

Third, translation tests accommodate measurement of PWK. Translation test scores do not have to be "right" or "wrong." When only two levels are recognized, we have "binary scoring," which is by definition insensitive to PWK. However, we do not have to limit ourselves to binary scoring. To distinguish PWK from other levels of word knowledge, we can adopt further levels of scoring, including "correct," "incorrect," and one or more intermediate levels. We refer to such systems as "multi-level scoring."

On the other hand, multilevel rating of open format L1 translation tests raises the question of how many "levels" should be recognized. Several researchers have adopted a three-level scale that includes an intermediate score for partially correct translations in addition to the "correct" and "incorrect" scores. A disadvantage of this approach is that all partially correct responses are "lumped together:" there is no differentiation between a translation that is "nearly correct" and one that is "nearly incorrect." Other authors have experimented with four- and five- level scales, in which PWK translations are assigned varying scores. An underlying problem has been the rationale for assigning differing PWK scores. For example, some rating systems are based on specific assumptions about the stages of the vocabulary acquisition process. One drawback of this approach is that the assumed sequence of stages may not be universally accepted. Another is that because each stage is characterized by a different form of production, the score may vary depending on the type of response. The difficulty of comparing two qualitatively different responses may make it impossible to verify that one response is more accurate than the other. (The literature review gives a specific example of this problem.)

In this paper, we employ four different rating scales to analyze experimental data gathered in a vocabulary acquisition study among English for Specific Purposes (ESP) students at a private university in central Taiwan. By comparing the impact of the scales on the experimental results, we answer the following research questions:

1. Are all partially correct L1 translations by the respondents equally accurate?
2. Does using multiple PWK levels affect the outcome of statistical analysis?

The remainder of the paper is as follows. First, we briefly review the literature on foreign language incidental vocabulary acquisition and the testing methodologies used to assess incidental acquisition in previous studies. We then discuss our research methodology. The last two sections present and discuss the experimental results.

II. LITERATURE REVIEW

What does it mean to "know" a word for reading purposes? Nation (1990) identifies seven diverse facets of reading vocabulary knowledge (Nation, p. 36):

- What the word looks like
- The grammatical patterns in which it occurs
- The words or word types that can be expected before or after it
- How frequently it occurs
- In what sort of discourse we can expect to meet this word
- What the word means
- What other words are associated with it

The types of assessment employed in studies of vocabulary and reading are equally diverse. We describe a few assessment categories below.

1. Tests of Passive Vocabulary Size

These include tests that assess knowledge of frequency-based word sets. Beglar and Hunt (1999) develop a test of a single word set, the most frequent 2,000 words. However, the most famous test of vocabulary size, Nation's Levels Test, in its original and updated versions (Nation, 1990; Schmitt, 2000) assesses knowledge of several word sets, including all words under the 10,000 frequency level and the University Word List. The Levels Test is available in multiple

formats, including matching and gap completion.

An alternate approach to measuring vocabulary size is to use a vocabulary checklist such as the Yes/No Test (Meara & Jones, 1990) This is an example of self-reported word knowledge, in which the subject identifies the words that (s)he thinks (s)he knows among the words presented on the test.

By definition, tests of passive vocabulary size measure the approximate size of a subject's vocabulary rather than the extent of partial word knowledge. Some authors have observed that the Levels Test is sensitive to partial word knowledge (Beglar, 2000). However, they clearly refer to the ability of the items to *detect* partial (i.e., imperfect) knowledge of the target vocabulary as opposed to *measuring its extent*.

2. Tests Custom Designed to Measure Knowledge of Smaller Word Sets

The goal is not to assess vocabulary size, but to determine which words in a given word set are known. Checklists based on self-reported word knowledge are commonly used for this purpose, for example when a researcher would like to eliminate from consideration target words that are known by too many subjects. One drawback of this test format is the temptation of subjects to over report their word knowledge (Anderson & Freebody, 1983; Waring, 2002). One way to circumvent this problem is to incorporate authentic-looking pseudowords in addition to the real words in the checklist. To discourage exaggeration, the subjects are informed of the presence of the pseudowords, as in (Nagy, et al., 1985).[2] Like tests of vocabulary size, this type of assessment is of limited use for detecting PWK gains.

Third are tests of vocabulary in a specific reading context. This type of test assesses knowledge of the interaction between the target word and its context. A good example is Webb (2007). His battery of 10 tests includes three that measure receptive knowledge of vocabulary in context: grammatical function, syntax, and word associations.

Webb used his elaborate test regimen to gain a comprehensive picture of vocabulary knowledge and how it changes with repeated exposure through reading. His work is a reaction to what he views as an excessive focus on word meaning: "… the majority of past research has equated gains in knowledge of meaning with acquisition" (Webb, 2007, p. 46). He views word knowledge as multifaceted, encompassing disparate aspects such as spelling, grammar and word associations. It follows that partial knowledge arises whenever one of these facets has not been fully developed.

This study adopts the narrower, meaning-focused notion of word knowledge, as stated in Swanborn and de Glopper (1999, p. 262): incidental vocabulary acquisition is "the incidental, as opposed to intentional, derivation and learning of new word meanings by subjects who are reading under reading circumstances that are familiar to them." We focus on meaning for two reasons. First, assessment of target word meanings (as understood by the subject) is the common denominator in most read-and-test studies. Some studies may evaluate other facets of word knowledge (e.g., understanding of a word's associations or how it is used in context), but almost all assess knowledge of word meaning. Second, there is a practical need for measures that enable us to observe and compare incremental gains in knowledge of meaning over time.

3. Tests of Word Meanings

Multiple choice assessment. Multiple choice assessment has been used to test vocabulary acquired through extensive reading. The first was a study by Saragi, Nation, and Meister (1978) of acquisition of *nadsat* (quasi-Russian slang words) by L1 readers of *A Clockwork Orange*, later replicated by Pitts, White, and Krashen (1989) among ESL students. Ferris (1988) (cited in Horst, 2000, Chapters 1 and 4) studied ESL students' acquisition of English target words from reading George Orwell's *Animal Farm*. In a well-known foreign language study, Horst, Cobb, and Meara (1998) investigated English vocabulary acquired by native Arabic speaking university students from reading a simplified version of Thomas Hardy's novel, *The Mayor of Casterbridge*.

Multiple choice items and multiple choice-like items have also been used in incidental vocabulary acquisition studies based on shorter texts. Zahar, Cobb, and Spada (2001) designed a test similar in format to Nation's Vocabulary Levels Test (Nation, 1990) to assess gains among French-speaking adolescent ESL students who had read a 2,381-word short story (*The Golden Fleece*). Day, et al. (1991) and Luppescu and Day (1993) assessed gains among Japanese high school and

[2] More recently, pseudowords have been employed as substitutes for authentic L2 target words in studies of incidental vocabulary acquisition such as (Reider, 2002; Waring & Takaki, 2003; Webb, 2007). One of the main goals is to ensure that the subjects start with zero word knowledge. (Another is to prevent subjects from consulting their dictionaries.) However, when the subjects already have partial or full knowledge of the original L2 words, there are subtle contradictions in the use of pseudoword substitutes. One is between the original L2 word, which is already present as an entry in the learner's mental lexicon, and the pseudoword, which is identical in meaning. The principle of contrast, as set forth in (Lockett & Shore, 2003), states that "No two words can mean exactly the same thing". The authors of this paper point out that a vital stage in a learner's acquisition of a target word is understanding how it differs in meaning from the words already in his/her mental lexicon. If the reading contexts all fit the original, authentic L2 word, the subject will gravitate towards this meaning. Meanwhile, the contrast principle will lead the subject to search for another meaning. The two forces will act in opposing directions.

university students who had read a shorter story, *The Mystery of the African Mask* (1,032 words).

Multiple choice tests are subject to various limitations. One is that they encourage guesswork (Horst, 2000). Waring and Takaki (2003) refer to multiple choice tests as "prompted meaning recognition" and point out that they are inherently easier than unprompted recall, e.g., open-format responses such as L1 translations, L2 definitions, or example sentences. Another problem is the insensitivity of multiple choice problems to partial gains. Some authors direct this criticism at multiple choice assessment (Horst & Meara, 1999). Others appear to direct it at binary response items in general (Horst, 2000; Nagy, et al., 1985; Webb, 2007).

Some multiple choice vocabulary tests circumvent this problem through imaginative use of distractors. In their work with American schoolchildren, Nagy, Herman, and Anderson extensively used multiple choice questions with five answer selections. One such test was used in a study of incidental acquisition among American schoolchildren in 3rd , 5th , and 7th grades (Nagy, Anderson, & Herman, 1987). Each item presented the target word and five choices: a correct definition, three distractors, and "I don't know." The distractors were definitions of real English words semantically related to and of the same part of speech as the target words.

Variants of this five-choice format have been used by these and other authors. A more elaborate implementation appears in Joe (1998), which employs five-choice tests in two versions: "easy" and "difficult." As in Nagy, et al. (1987), each item consisted of the target word followed by a correct definition, three distractors, and an "I don't know" option. In the easy version, the correct choice was a general definition of the target word, and each of the distractors was "semantically different" from the target word, but "still within the same topic domain." The distractors did not necessarily have the same part of speech as the target word. In the difficult test, the distractors were "semantically close" to the target word and had the same part of speech, making selection of the correct response more challenging.

Self-reported word knowledge. In the 1990s, researchers began to experiment with self-reported word knowledge, in which the subject is presented with a list of words and rates his/her familiarity with each. An example is the Yes-No vocabulary test (Meara & Jones, 1990), in which the subject marks each word as known or unknown. Most studies of incidental vocabulary acquisition that employ self-reported knowledge, however, allow for at least three levels of word knowledge, including one or more intermediate levels representing PWK. In their experimental model of vocabulary acquisition, Horst and Meara (1999) give their subjects three choices, including one intermediate response ("I'm not sure whether I know this word."). In a pretest/posttest study, Watts (2002) allows for four levels of self-reported knowledge, including two levels of PWK.

An interesting variant on self-reported word knowledge is the vocabulary knowledge scale (VKS) of Wesche and Paribakht (1996), a hybrid, five-level rating scheme in which levels 1-2 are reserved for self-reported word knowledge and levels 3-5 are for tentative L1 translations, confident L1 translations, and example sentences, respectively. Schematically, we can represent this scale as an iceberg, as shown in Figure 2. The top levels are visible manifestations of the subject's word knowledge, and portrude above the surface of the water. The lower levels, which presumably represent earlier stages in the accumulation of word knowledge, are observable to the rater only through the subject's self-reported knowledge, and rest below the surface.

Experimenters have justified their use of self-reported knowledge and hybrid scales such as the vocabulary knowledge scale on the ground that they are sensitive to partial word knowledge. However, both types of assessment have drawbacks. One is the temptation of subjects to over report their word knowledge (Anderson & Freebody, 1983; Waring, 2002). Another is the apparently arbitrary equating of numerical scores with qualitatively different responses, such as self-reported familiarity and L1 translations, in scales such as the vocabulary knowledge scale (Waring, 2002).

Definitions and Translations. A third type of assessment is open-format (unprompted), user-supplied definitions. This is a broad family of item responses that includes both target language definitions and L1 translations, the focus of this paper. Examples of the former approach include Hermann's (2003) study of gains among adult ESL readers of George Orwell's *Animal Farm* and a study by Swanborn and de Glopper (2002) of English vocabulary gains among Dutch sixth grade readers of an L1 grade-level informative text. Both

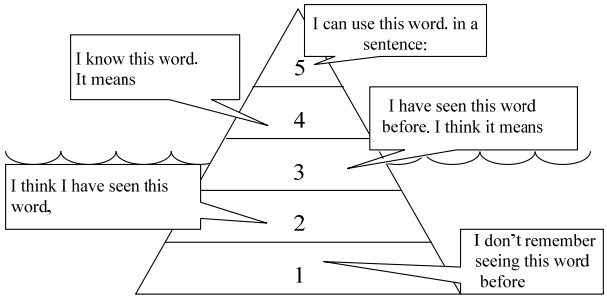


Fig. 2. Iceberg representation of the VKS of Wesche & Paribakht (1996)

studies employ a similar four-point scale, ranging from 0 (no word knowledge) to 3 (full word knowledge). Beyond the rating scale, however, there are substantial differences between the tests employed by the two authors. Hermann's test was implemented as a modified cloze test, in which each sentence was to be completed by a "periphrastic" definition of its associated target word. Hermann employed this format because the sentence contexts in the test "severely restricted the range of acceptable definitions subjects could provide," for example by specifying the target word sense (when more than one word sense was possible) and part of speech. By comparison, responses in Swanborn and de Glopper's test were unrestricted. Subjects who could not provide a definition could (for lesser credit) write an example sentence using the target word.

When the subject's ability to provide target language definitions is limited by poor L2 ability, an L1 translation test is preferable to an L2 definition test. The sensitivity of these tests depends on the scoring procedure. Binary scoring of translations does not recognize PWK of word meaning.

In addition to correct translations (FWK) and incorrect translations, some tests recognize one level of PWK. Laufer (2003) conducted an experiment comparing vocabulary gains among Israeli university EFL learners resulting from two different treatments: 1. sentence writing and 2. reading of a 621-word newspaper article with Hebrew glosses of the target words. In the posttest, the subjects supplied L1 translations and/or definitions of the target words. Correct responses received a score of 1.0, incorrect responses, 0.0, and "semantically approximate explanations or translations," 0.5. A similar posttest was employed in a study of vocabulary gains among 95 undergraduate foreign language German learners by Rott (1999). For the study, Rott custom wrote 72 short paragraphs containing 12 target words. During the readings, her subjects met each word 2, 4, or 6 times. After the treatment, they completed an L1 translation test. The translation responses were rated as unknown (0), partially known (1), or fully known (2). A third example of a three-level translation posttest is described in (Waring & Takaki, 2003), a study of vocabulary acquisition from a 5872-word short story in a graded reader among 15 Japanese female English learners. After reading the story, the subjects were given a translation test for knowledge of 25 pseudowords which had been inserted into the story as synonyms for common English words. The responses were rated as follows: 0 for an incorrect translation, 1 for a correct translation, and .5 for PWK (a "similar word").

Because the subject may achieve only PWK within the time frame of a study, tests that recognize only full word knowledge (FWK) risk underestimating incidental vocabulary gains, as pointed out by Swanborn and d Glopper (1999). Over time, researchers have come to recognize the importance of PWK assessment, and there has been a corresponding drift from binary scoring (in which responses are strictly classified as "correct" or "incorrect") towards multi-level scoring (in which at least one intermediate level of word knowledge is recognized).

4. The Role of Translation in the Formation of the L2 Mental Lexicon

Psycholinguistic research indicates that the L1 translation process may play a crucial role in the productive use of L2 vocabulary. Learners, particularly those who are less fluent, tend to access L2 words by first retrieving their L1 translations (Kroll & Curley, 1988; Kroll & Stewart, 1990).

The difference between L1 processing and target language processing may be more pronounced for Chinese-speaking EFL learners than for learners whose L1 is more closely related to English. For example, closely related languages are likely to contain form-similar cognates (e.g., the English word "music" and the Spanish word "musica"). It is known that adult bilinguals translate noun cognates faster in both directions (L1→L2 and L2→L1) than they do non-cognates (Kroll & de Groot, 1997), a phenomenon is known as the Cognate Facilitation Effect (CFE).

The CFE illustrates the phenomenon of "positive transfer" among learners' whose L1 is closely related to the target language (Gass, 1983). The counterpart to positive transfer is "negative transfer" (Gass, 1983), which impedes the efforts of learners whose L1 is more distantly related to the target language. Negative transfer – alternately referred to as "language interference" - helps to explain the relative difficulty of English vocabulary acquisition among Chinese-speaking EFL learners.

In L2 vocabulary acquisition, there is a gradual liberation from L1 language interference, as shown in the developmental model presented in (Jiang, 2000). This is a slower process for learners whose L1 is distant from L2. Another problem that affects learners in Taiwan is an overdependence on L2 classes. "Classroom L2 learners often lack sufficient, highly contextualized input in the target language. This often makes it extremely difficult, if not impossible, for an L2 learner to extract and create semantic, syntactic, and morphological specifications about a word and integrate such information into the lexical entry of that word." (Jiang, p. 52)

III. METHOD

This paper focuses on how rating precision impacts the results obtained from translation–based assessment of word

knowledge. We base our ideas on the analysis of L1 translation data obtained from pretests and posttests of words from the Academic Word List (AWL) of Coxhead (2000). In this section, we provide an overview of the data collection and scoring. We then describe our strategy for analyzing the data.

1. Participants

The participants were 64 undergraduate mechanical engineering majors at a private university in Central Taiwan. All were upperclassmen (juniors or seniors) enrolled in a one-semester, 3 credit ESP course. The course was offered as a new elective in the mechanical engineering department, and attracted an enrollment of 109 students, not including students who dropped out during the first few weeks of class. A total of 99 students completed the pretest and 78 took the posttest. We selected as participants the 64 students who completed both the pretest and the posttest.

2. Materials

The materials included required reading and listening, a vocabulary list, a pretest, and a posttest. All were based on course materials from two textbooks, *Oxford English for electrical and mechanical engineering* (Glendinning, E. & Glendinning, N., 1995) and *Basic English for computing* (Glendinning & McEwan, 2003).

The reading and listening materials. The students were required to read and complete the written exercises for the first 15 units of each textbook. The listening exercises were done in class, with the help of the teacher. All work in the textbooks is in the form of numbered "tasks." This made it easy for the instructor to make homework assignments and keep track of what was done during class hours.

The vocabulary list. We derived the vocabulary list from the reading and listening materials as follows.

First, we created a mini-corpus consisting of 60 electronic files. Half of these files contained the text from the first 15 units of each textbook and the other half contained transcripts for the corresponding audiotapes.

Next, we ran a custom-developed 'C' program that automatically searched through the text files and counted the occurrences of the words in each AWL word family. For each of the 60 input files, the program generated 570 AWL frequency counts. The word counts allowed us to determine how many AWL words were used in each book and the number of occurrences of each of them.

With the help of the frequency counts, we then created a list that included 54 target words that occurred in the textbooks and an additional 36 AWL control words that did not. The target words were selected according to the following criteria:

A. There was no alternate word sense that the students were likely to know. For example, the word "medium" (sublist 9) has both a subtechnical sense (e.g., a *medium* of instruction) and a conversational sense (a class of *medium* difficulty).

B. They were not overly familiar to students, according to the authors' experience. Words such as "computer" (sublist 2), "data" (sublist 1), and "tape" (sublist 6) were disqualified.

C. They were divided equally between low, mid, and high frequency words. The target words ranged in frequency from 1 to 24 occurrences in the mini-corpus. There were about equal numbers of low frequency (2-4 occurrences), mid-frequency (5-8 occurrences) and high-frequency (9 or more) words.

D. They were distributed over as many sublists as possible. There were at least four words from each of the first eight AWL sublists.

The purpose of the control words was to compare the subjects' knowledge of the subtechnical words in the textbooks with their knowledge of the AWL as a whole. The first author, an experienced technical editor, chose words that she had frequently encountered in engineering literature. The control words were selected so that their distribution over the AWL sublists would mirror the target word distribution.

Appendix 1 lists the target words. An asterisk (*) indicates a control word.

The vocabulary pretest and posttest. Both the pretest and posttest were open-format translation tests based on the words in the vocabulary list. In each test item, an English word was presented and the subject was asked to provide a translation into Chinese. A cue (a short phrase containing the word) was provided with each word to help the subjects determine the word sense and part of speech of the target words as they were used in the course textbooks.

Table 1 shows how the target words were allocated among the subjects during the pretest and posttest. Rather than presenting all 90 words to each subject on each of the two tests, we adopted an alternate procedure that was less tiring for the students and required less class time. During each test, the students were divided into nine groups and the subjects in each group were given a subset of 40 words. The word choice was randomized so that the tests for all the groups spanned the whole 90-word list. Within each group, the order of the words as they appeared on the test paper was also randomized to compensate for the effects of fatigue.

3. Procedure

On the first day of class, the students were given a schedule that listed the tasks in each textbook that were to be

Table 1. Allocation of the words among subjects during the pretest and posttest

Group	Number of Subjects	Words Tested in the Pretest	Words Tested in the Posttest
1	8	1-10, 61-90	21-60
2	9	1-20, 71-90	31-70
3	1	1-30, 81-90	41-80
4	9	1-40	51-90
5	5	11-50	1-10, 61-90
6	8	21-60	1-20, 71-90
7	9	31-70	1-30, 81-90
8	6	41-80	1-40
9	9	51-90	11-50
9 groups	64 subjects	40 words per subject	40 words per subject

completed in class and as homework. If an in-class task was not completed on the scheduled day, it became a homework assignment. At no point were the students told about our research. No copy of the vocabulary list was provided to the students. The students encountered the target vocabulary during classroom listening and during their individual work in the textbook. During class sessions, the instructor offered no verbal explanations of the target word meanings.

We administered the pretest on the second day of class so the results would not be affected by exposure to the materials during the course. We explained to the students that the purpose of the test was to gather information about their vocabulary knowledge. All test papers were individually numbered and collected. The posttest was administered on the day of the final exam, before the final exam. Again, the test papers were numbered and collected. The pretest and the posttest each took about 20 minutes of class time.

4. Rating of the Pretest and Posttest

After data collection, the pretest and posttest translations were analyzed. For each of the 90 words, we listed every translation that had appeared in at least once in either the pretest or posttest papers. The result was a master list of translations for each word. The master translation lists were developed into rating sheets. The 90 rating sheets were printed out and given to two bilingual raters. The raters independently assigned scores ranging from 0 (inaccurate) to 5 (accurate) to each word translation.

During the rating process, we monitored the raters' work. When the scores for a translation diverged too widely, we asked the raters to enter into discussion. However, when the raters were not able to reach agreement, we left the score difference "as is."

After completion of the translation scoring, we averaged the scores that the two raters assigned to each translation. We used the averages to grade the pretests and posttests.

5. Analysis Strategy

As the students completed their work in the course textbooks, they acquired knowledge of some of the target words. This type of acquisition is very similar to incidental acquisition,[3] and is therefore typified by PWK gains. Our goal in this paper is to compare the effect of rating precision on our ability to detect these small gains.

During the rating process, the raters studied each translation and assigned an integer score from 0 to 5. The scores of the two raters were averaged, introducing five new, non-integer scores (.5, 1.5, 2.5, 3.5, and 4.5). All together, there were 11 scoring levels. Therefore, we refer to the resulting numbers as eleven-level scoring.

From the eleven-level scores assigned by the raters, we can derive scores of lower precision as follows:

Six-level scoring. We obtain six-level scoring by rounding off the original scores to the nearest integer. Thus, a .5-point score is rounded to 1 point, a 1.5-point score is rounded to 2 points, and so on.

Three-level scoring. If the original score is more than or equal to zero and less than 5/3, it is assigned a three-level score of 0. If it is more than or equal to 5/3 and less than 10/3, it is assigned a three-level score of 1. If it is more than or equal to 10/3 and less than or equal to 5, it is assigned a three-level score of 2.

Binary scoring. If the original score is less than or equal to 2.5, it is assigned a binary score of 0. Otherwise, the binary score is 1.

To compare the effect of scoring at different levels of precision, we compare the pretest and posttest data for each word using the Mann-Whitney *U* test. Because there are four levels of precision and a total of 90 words, this means running the *U* test 360 times. We take the results obtained from the original, eleven-level scores as a golden standard. Through comparison, we determine the rate of errors in the analysis based on data of lower precision. We look for the following types of errors:

False positives. Based on the lower precision data, the *U* test flags a word for significant change between the pretest and the posttest. Based on the eleven-level data, the *U* test

[3] The students did their work either at home or in class, under crowded conditions. Therefore, it was not possible for us to monitor or control how they did their work. The students were unaware that we were targeting particular words for a posttest. On the other hand, some of them surely consulted their dictionaries during assignments. Therefore, this was not (strictly speaking) a study of incidental acquisition.

determines that the pretest / posttest change for this word is not significant.

False negatives. Based on the eleven-level data, the *U* test flags a word for significant change between the pretest and the posttest. Based on the lower precision data, the *U* test determines that the pretest / posttest change for this word is not significant.

IV. RESULTS AND DISCUSSION

1. Research Question 1: Are All Partially Correct L1 Translations Equally Accurate?

In this section, we demonstrate that the answer to this question is "no". We look at this question from the perspective of the two raters. First, we examine the distribution of the PWK scores within the scoring range. We then look at the rating assignments for translations of one of the 90 target words.

Distribution of PWK scores. Collectively, there were 1285 different target word translations among the pretest and posttest responses. The number of translations varied from word to word, from a minimum of 5 for the word "similar" to a maximum of 27 for the words "detect" and "principle." The score for a translation (ranging from 0 to 5) was the average of the scores assigned by the two raters. There were thus 11 levels: 0 for a point value of 0.0, 1 for a point value of .5, 2 for a point value of 1.0, and so on. From these point assignments, we can derive equivalent ratings under a binary, three-level, and six-level scoring system.

Table 2 gives breakdowns of the translations by score for each of the four rating precisions, shown in histogram form in Figure 3. As the scoring precision rises, the proportion of the translations classified as PWK also rises, from 0% for binary scoring to 5.7% (=73/1285) for three-level scoring, to 18.0% (=(72+77+42+40)/1285) for six-level scoring, and finally, 19.6% (=(52+20+46+31+20+22+18+22+21)/1285) for eleven-level scoring. By far, the majority of translations were inaccurate. At the eleven-level precision, about 75% of the translations (969 out of 1285) received a score of 0. Just under 5% (64) received the full 5 points, meaning that for some of the 90 target words, no respondent was able to give a fully accurate translation.

An alternate way to look at the distribution of PWK scoring is to consider the proportion of test responses that constituted PWK. Table 3 gives breakdowns of the pretest translation responses by level at each of the four rating precisions, which are shown in histogram form in Figure 4. Clearly, of the 2660 responses (=64 subjects×40 items/subject), only a minority constitute PWK. In terms of the relationship between scoring precision and PWK assignments, Table 3 is similar to Table 2. As the scoring precision rises, so does the percentage of responses that are considered neither inaccurate nor FWK. The percentage of responses in the middle range rises from 0% for binary scoring, to 3.3% (=88/2660) for three-level scoring, to 14.4% (=(68+94+56+164)/2660) for six-level scoring, to 17.9% (=(58+10+62+32+37+19+24+140+94)/2660) for eleven-level scoring.

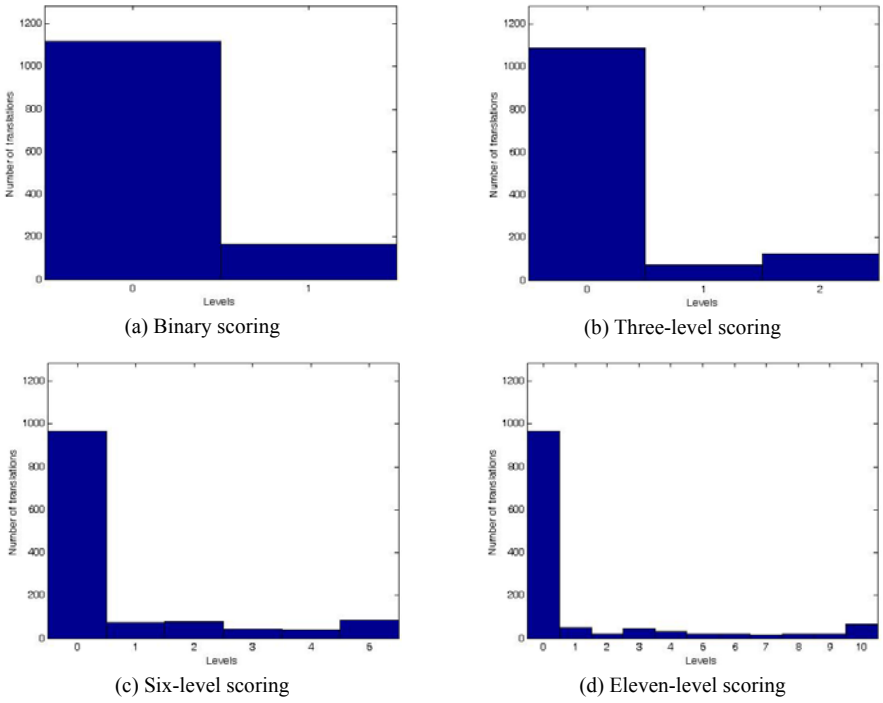
Conclusion. From the experience of the raters, we conclude that the answer to research question 1 is "no." Both the individual Chinese translations of the target words and the pretest score assignments were spread throughout the scoring range. With eleven-level scoring, nearly 20% of translations and pretest scores were judged as PWK. The translations of the example target word illustrate (in a subjective way) how two PWK responses might be rated "unequal." Tables 2 and 3 illustrate that lower scoring precision has the effect of reducing the number of PWK assignments.

2. Research Question 2: Does Using More PWK Levels Affect the Statistical Outcome?

In this section, we use error analysis to show that the answer to this question is "yes". First, we compare the pretest and posttest data at full precision (eleven-level scoring) to determine for which target words the subjects' knowledge improved. We repeat this analysis using binary, three-level, and six-level score data. Using the eleven-level outcomes as a standard, we show that binary and three-level scoring leads to errors – that is, to different results from six- and eleven-level scoring.

Table 2. Breakdowns of the 1285 Chinese translations by score

Precision	Scoring level										
	0	1	2	3	4	5	6	7	8	9	10
Binary	1118	167									
Three-level	1087	73	125								
Six-level	969	72	77	42	40	85					
Eleven-level	969	52	20	46	31	20	22	18	22	21	64



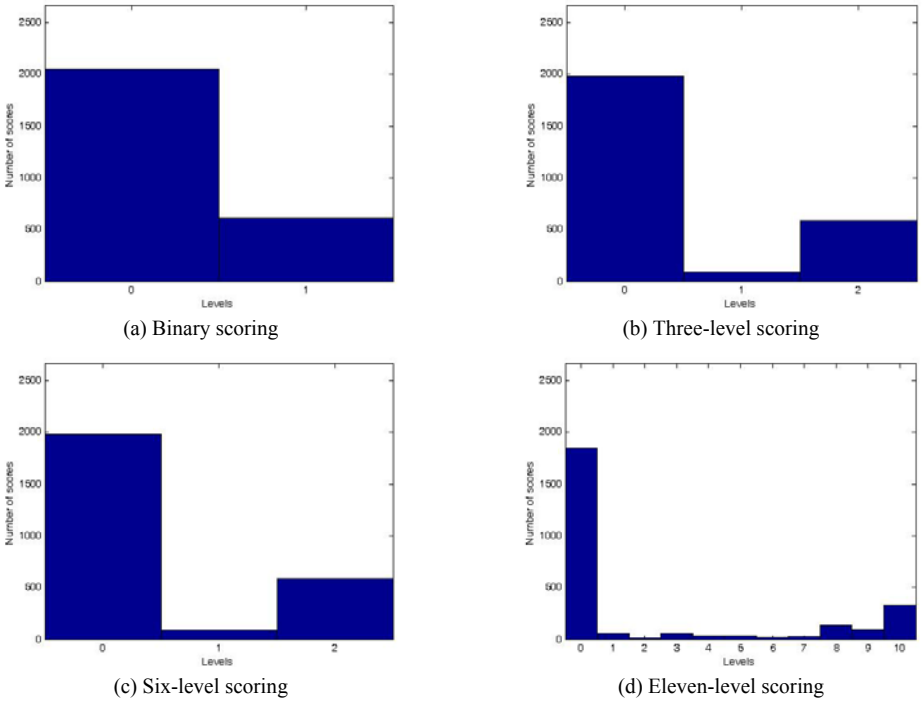
(a) Binary scoring (b) Three-level scoring

(c) Six-level scoring (d) Eleven-level scoring

Fig. 3. Histogram breakdowns of translations by level

Table 3. Breakdowns of the pretest scores by level

Precision	Scoring level										
	0	1	2	3	4	5	6	7	8	9	10
Binary	2050	610									
Three-level	1981	88	591								
Six-level	1851	68	94	56	164	427					
Eleven-level	1851	58	10	62	32	37	19	24	140	94	333



(a) Binary scoring (b) Three-level scoring

(c) Six-level scoring (d) Eleven-level scoring

Fig. 4. Histogram breakdowns of pretest scores by level

Change in word knowledge between the pretest and the posttest. In this study, as in numerous studies of incidental vocabulary acquisition among foreign language learners, the improvements in word knowledge were very small. Of the 90 words tested, the subjects changed significantly in their knowledge of only 11 words, listed in Table 4. Column 1 gives the target word; column 2, the pretest average score and the number of subjects who were tested on the word in the pretest; column 3, the posttest average score and the number of subjects who were tested on the word in the posttest; column 4, the difference between the pretest and posttest averages; and column 5, the *p* values obtained from comparison of the pretest and posttest scores. Because the translation scores are not normally distributed, we compared the pretest and posttest data using the Mann-Whitney *U* test instead of a *t*-test.

The magnitude of the pretest/posttest average score change for the words in Table 5 varies considerably. The lowest change (-0.31) corresponds to the only word for which knowledge decreased significantly – strangely enough, the word "significant." The highest change was for the word "impact," the only word for which knowledge increased by more than 2.5 points (half of the score range). On average, knowledge of the words in Table 1 changed by only 1.19 points.

False positive results in two- and three-level scoring. Tables 4 and 6 compare the results obtained from two-level, three-level, six-level, and ten-level scoring of the translation data. Table 4 shows false positives – that is, words for which the subjects' vocabulary knowledge was incorrectly identified as undergoing significant change between the pretest and the posttest. Column 1 shows the target word. Columns 2 and 3 show the average of the pretest and posttest scores, calculated on the basis of the eleven-level scores (which are taken as a "golden standard" of comparison), and the numbers of subjects who were tested on the word during the pretest and the posttest. Columns 4 to 7 show the *p* values obtained from the Mann-Whitney *U* test. The *p* value for each false positive is underlined and appears in boldface print.

The maximum number of false positives (3) occurs with binary scoring. However, three-level scoring also results in two false positives. As the number of scoring levels increases, the *p* values either increase or remain constant. For each of the target words in the table, there is a precision threshold. When the number of scoring levels falls below this threshold, the Mann-Whitney *U* test yields a false positive, but at or above the threshold, the *U* test results are negative. The thresholds for the target words "available" and "expand" are at either 4 or 5 levels. For the target word "construct," on the other hand, the threshold is at 3 levels.

By comparing the histograms of the translation data for each wrongly classified target word at each of the four levels of scoring precision, we can understand how scoring precision

Table 4. False positive results from the Mann-Whitney *U* Test at various scoring levels

Target word	Pretest avg. (N1)	Posttest avg. (N2)	2 lev	3 lev	5 lev	10 lev
			p	*p*	*p*	*p*
available	0.11 (27)	0.86 (28)	**.024**	**.024**	.114	.114
construct	0.46 (27)	1.36 (28)	**.027**	.150	.098	.098
expand	2.13 (32)	1.07 (27)	**.017**	**.026**	.063	.063
Total false positives			**3**	**2**	**0**	---------

Table 5. Target words for which knowledge changed significantly in the posttest

Target word	Pretest avg. (N1)	Posttest avg. (N2)	Change	*P*
consist	0.74 (23)	2.63 (32)	1.89	.002
corresponding	0.21 (24)	1.16 (32)	0.95	.033
contact	1.27 (22)	2.66 (32)	1.39	.024
estimate	0.44 (27)	2.09 (28)	1.65	.008
impact	2.22 (27)	4.82 (28)	2.6	~0
method	2.06 (24)	3.25 (32)	1.19	.039
plus	1.66 (22)	3.13 (32)	1.47	.041
reject	1.20 (27)	2.41 (28)	1.21	.030
resource	0.70 (27)	1.54 (28)	0.84	.041
significant (*)	1.14 (33)	0.83 (24)	-0.31	.010
survey	0.00 (32)	0.22 (23)	0.22	.039
Total words for which knowledge changed significantly			**11**	

Note: (*) Significant decrease in word knowledge.

Table 6. False negatives from the *U* Test at various levels of scoring precision

Target word	Pretest avg (N1)	Posttest avg (N2)	2 lev	3 lev	5 lev	10 lev
			p	*p*	*p*	*p*
corresponding	0.21 (24)	1.16 (32)	.027	**.051**	.033	.033
method	2.06 (24)	3.25 (32)	**.052**	**.052**	.039	.039
reject	1.20 (27)	2.41 (28)	**.070**	.031	.029	.030
resource	0.70 (27)	1.54 (28)	**.137**	**.137**	.041	.041
survey	0.00 (32)	0.22 (23)	**.235**	**.235**	.039	.039
Total false negatives			**4**	**4**	**0**	--------

affects the outcome of the *U* test. In each of Figures 5-10, the histograms are arranged in increasing order of precision, from binary scoring (parts 6(a), 7(a), and so on) to eleven-level scoring (parts 6(d), 7(d), and so on). Because the number of pretest and posttest subjects varies from word to word, we have normalized the histogram data to a scale of 100, so that the bars in the histograms represent percentages. To make the graphs easier to read, we have adopted integer indices for the eleven levels in the eleven-level histograms, so that level 1 corresponds to a score of .5, level 2 corresponds to a score of 1.0, and so on.

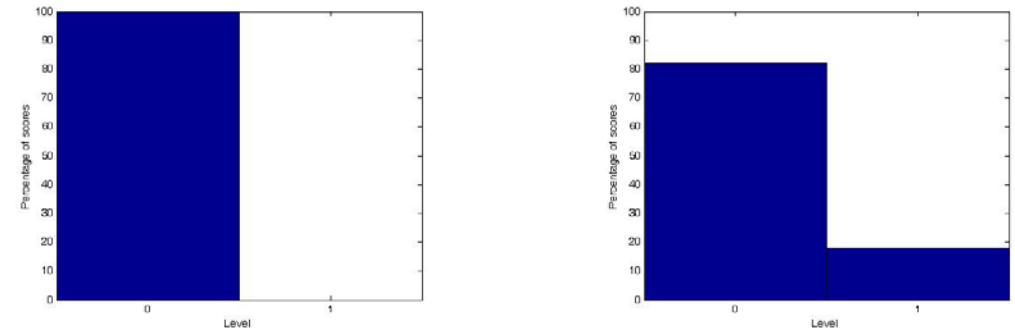
Figures 5 and 6 show histograms for two of the target words in Table 6. Figure 5 is for the target word "available." Subject knowledge of this word was extremely low. The average score increased slightly from .11 to .86 points at the posttest, however this change was not significant. A glance at part 5(d) shows that in the pretest and posttest, a few of .the subjects earned 1.5 points for two different partially correct translations (提供, meaning "provide," and 允許, meaning "allow"). Under both binary and three-level scoring, these translations earned no credit, making the pretest and posttest data more polarized and leading to the false positive result. The appearance of these PWK ratings in the six- and eleven-level histograms has the effect of spreading out the "zero" data from the binary histogram. In the process, it makes the pretest and posttest outcomes more closely resemble one another. Hence, at the six-level precision, the improvement in translation scores becomes insignificant.

Figure 6 shows histograms for the word "expand," one of the few target words for which the translation scores decline in the posttest. The average translation score decreases by more than one point, from 2.13 in the pretest to 1.07 in the posttest. However, because of the manner in which the scores are distributed, the difference is not significant. As in Figure 3, the cause is PWK translation scores. In this case, several subjects chose one of three 3.5-point translations (散開, meaning "spread," 展開, meaning "spread out," and 擴張, meaning "extend"), which appear in the middle of the six- and eleven-level ranges in Figures 6(c) and 6(d), but at the top end of the two- and three-level ranges in Figures 6(a) and 6(b). Because these same PWK scores are present in both the pretest and the posttest, they have the effect of making the pretest and posttest scores more similar. Once again, when we reach six-level precision, the change in translation scores becomes insignificant.

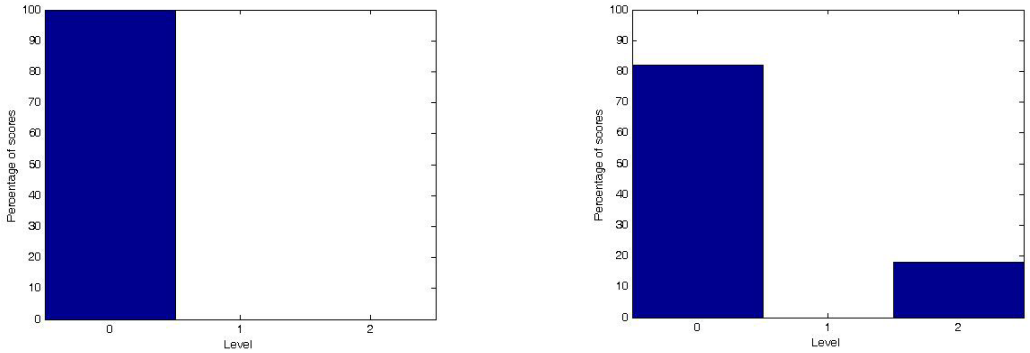
False negatives in two- and three-level scoring. Table 6 shows false negatives – that is, words for which the change in the subjects' vocabulary knowledge between the pretest and the posttest was incorrectly labeled insignificant. As in Table 4, column 1 shows the target word, columns 2 and 3 show the averages of the pretest and posttest scores and the numbers of respondents who were tested on the word, and columns 4 to 7 show the *p* values obtained from the Mann-Whitney *U* Test at each of the four precision levels. The *p* value for to each false negative is underlined and boldfaced.

Figure 7 shows histograms for the first target word in Table 6, "corresponding." The *U* test results for this word are anomalous: there is no clear cutoff point at which the false negatives disappear. Binary, six-level, and eleven-level scoring all yield positive results (i.e., a determination that the posttest translations are significantly better than the pretest translations), but three-level scoring yields a false negative. From the histograms in Figures 7(b)-7(d), it is not clear why the pretest/posttest differences at the six- and eleven-level precision are more "significant" than those shown at the three-level precision. We attribute this situation to the use of the Mann-Whitney *U* test, which is based not on the absolute size of the data, but on their relative size (rankings). For this reason, the *U* test occasionally yields counterintuitive results.

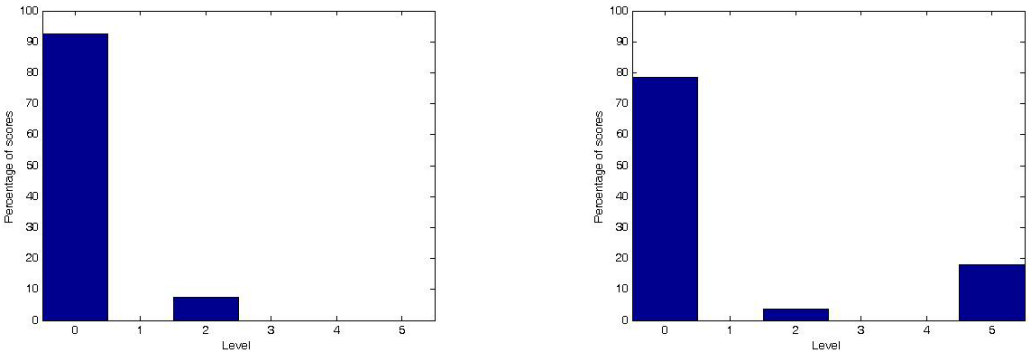
Figure 8 shows histograms for "reject," the third target word in Table 6. In Figure 8(a), there is a modest improvement between the pretest and the posttest: the number of 0 scores decreases, while the number of 1 scores increases. However, the change does not reach significance. In Figure 8(b), the posttest scores begin to look different: two of the subjects who earned zeros under binary rating now have a PWK score of 1.



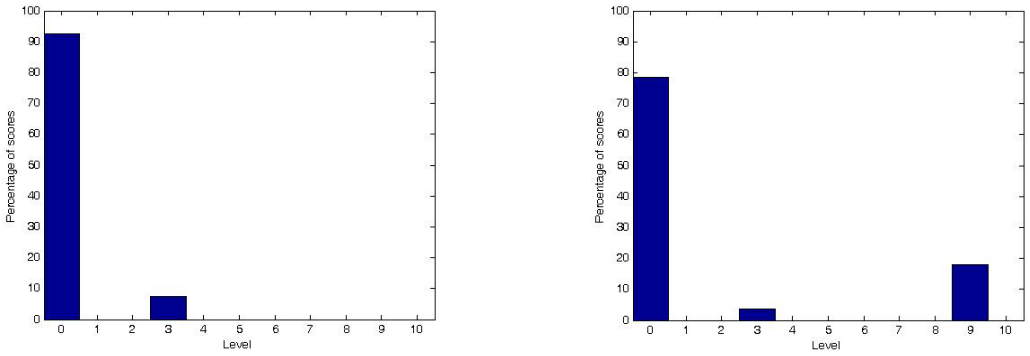
(a) Binary pretest and posttest scores (p = 0.24)



(b) Three-level pretest and posttest scores (p = .024)

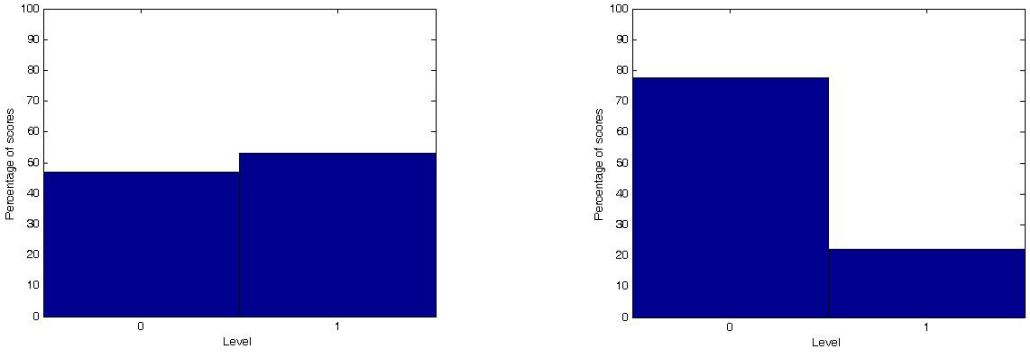


(c) Six-level pretest and posttest scores (p = .114)

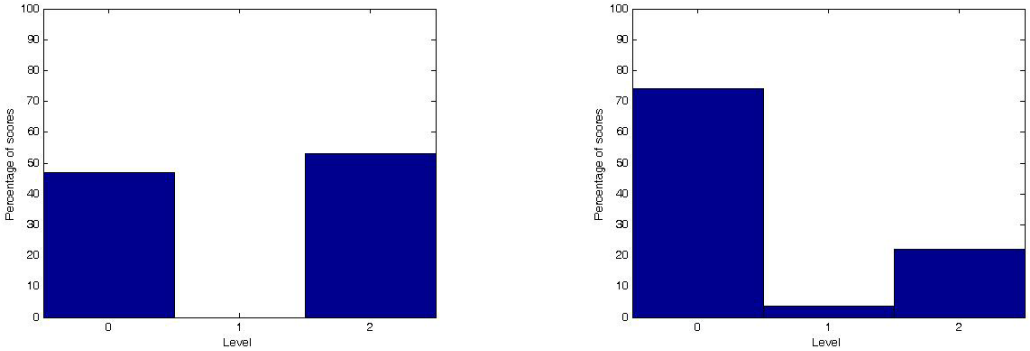


(d) Eleven-level pretest and posttest scores (p = .114)

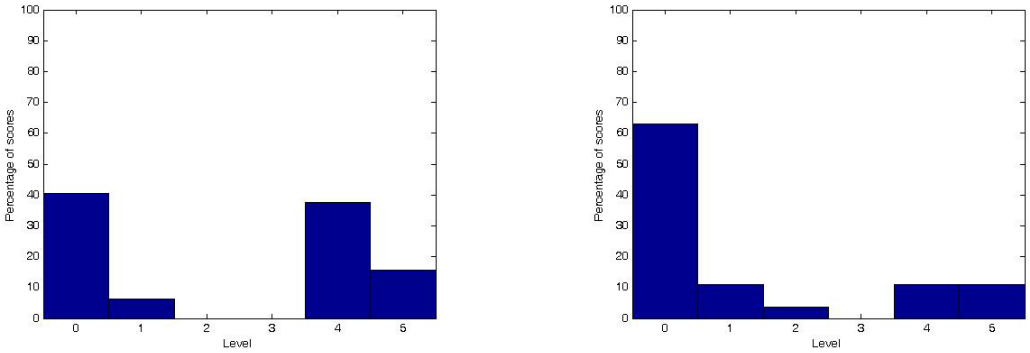
Fig. 5. Histograms for the target word "available."



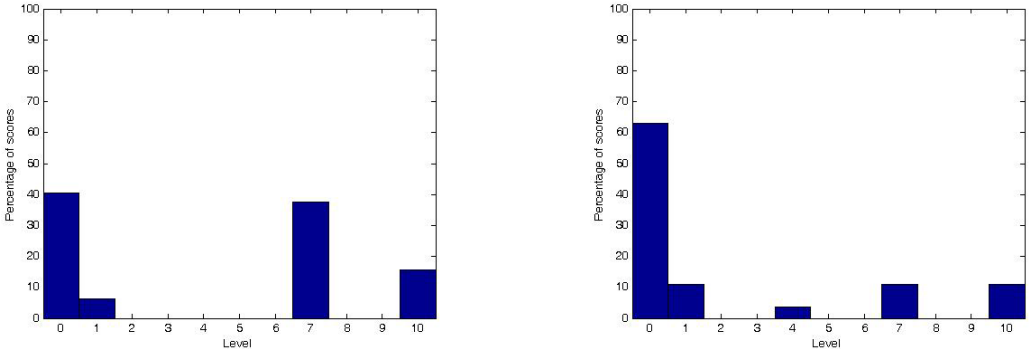
(a) Binary pretest and posttest scores (p = .017)



(b) Three-level pretest and posttest scores (p = .026)

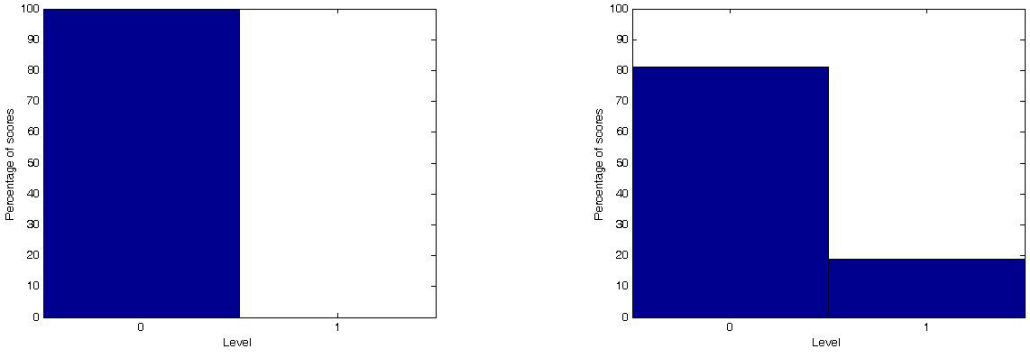


(c) Six-level pretest and posttest scores (p = .063)

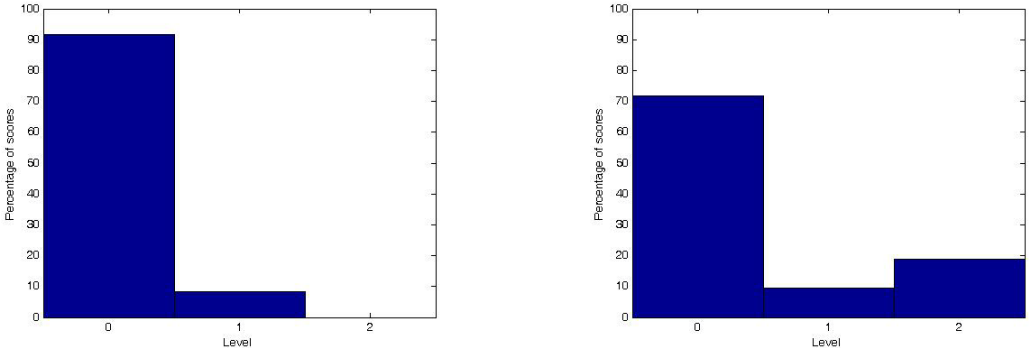


(d) Eleven-level pretest and posttest scores (p = .063)

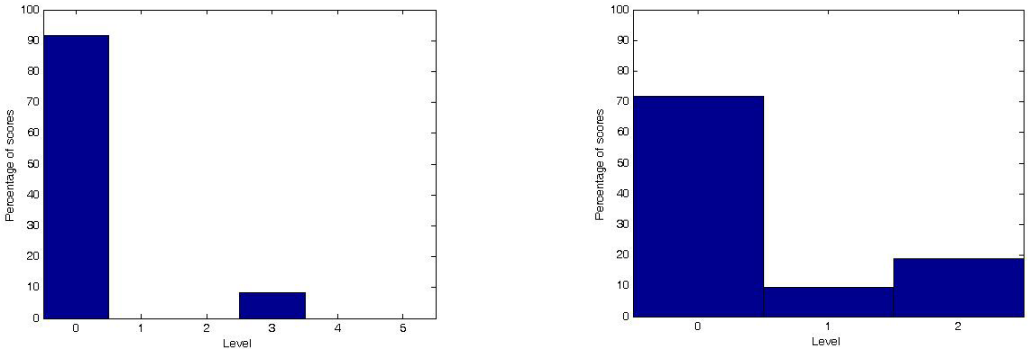
Fig. 6. Histograms for the target word "expand."



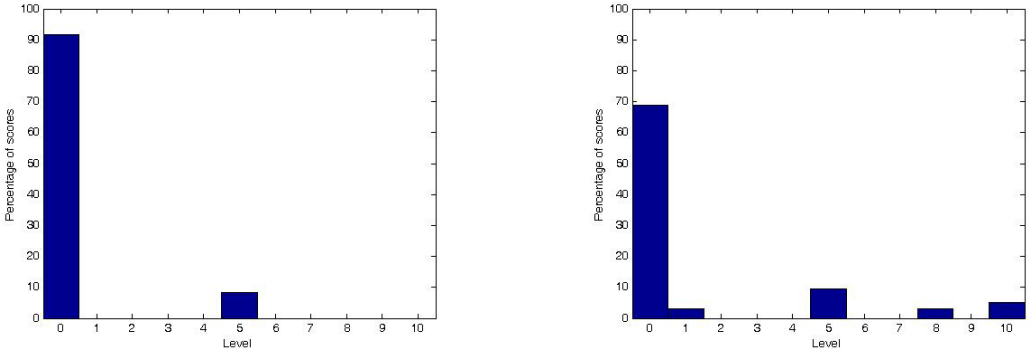
(a) Binary pretest and posttest scores (p = .027)



(b) Three-level pretest and posttest scores (p = .051)

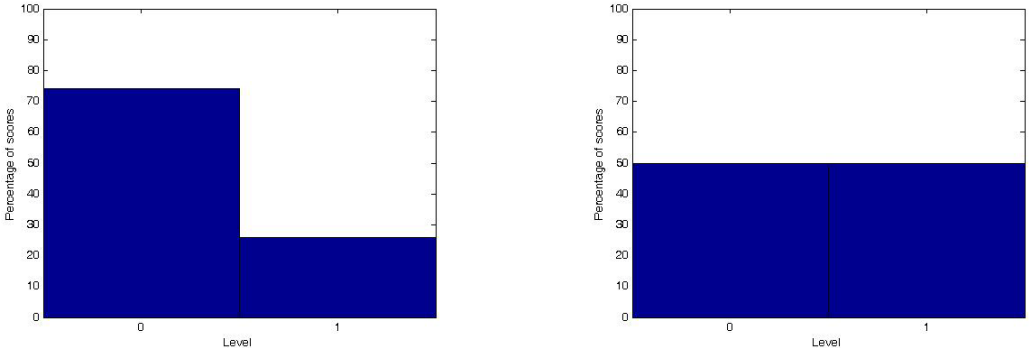


(c). Six-level pretest and posttest scores (p = .033)

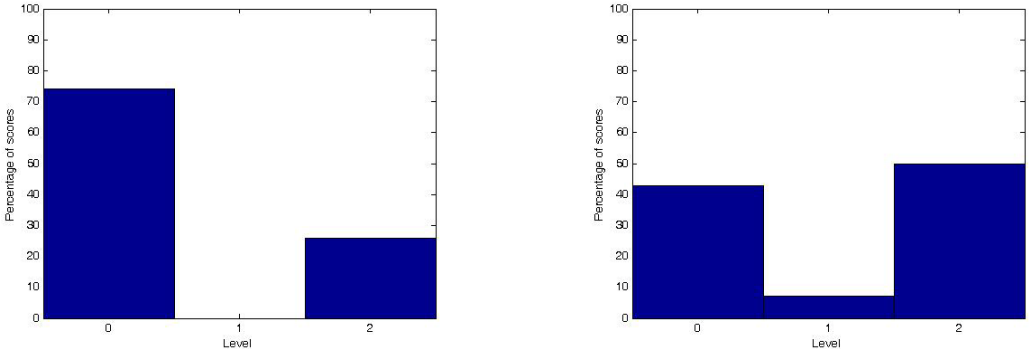


(d) Eleven-level pretest and posttest scores (p = .033)

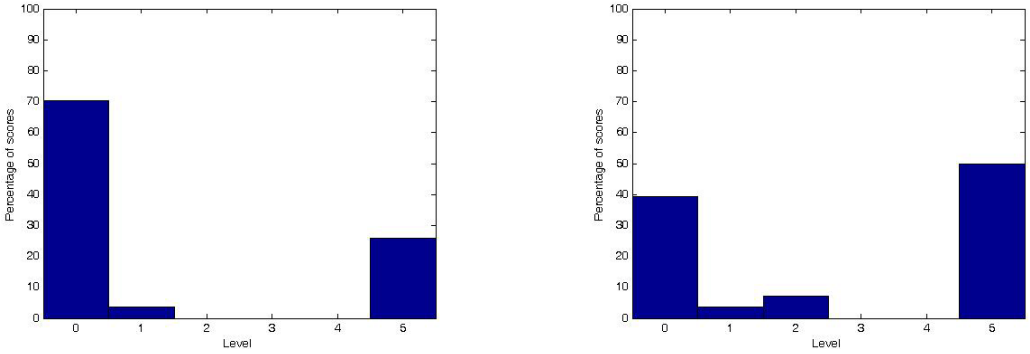
Fig. 7. Histograms for the target word "corresponding."



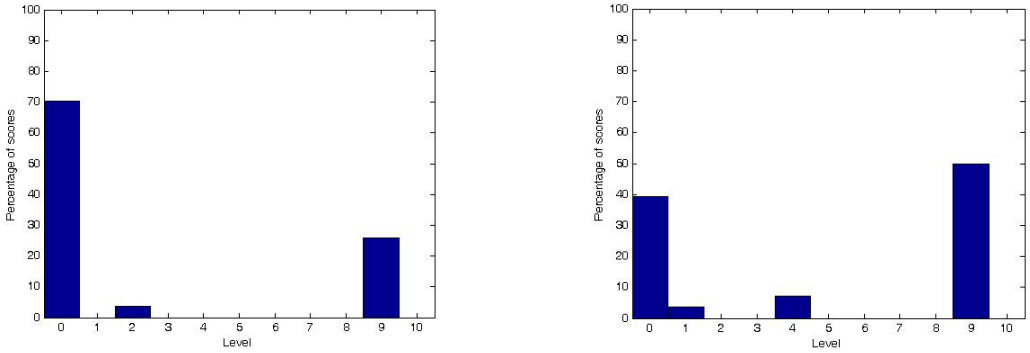
(a) Binary pretest and posttest scores (p = .070)



(b) Three-level pretest and posttest scores (p = .031)

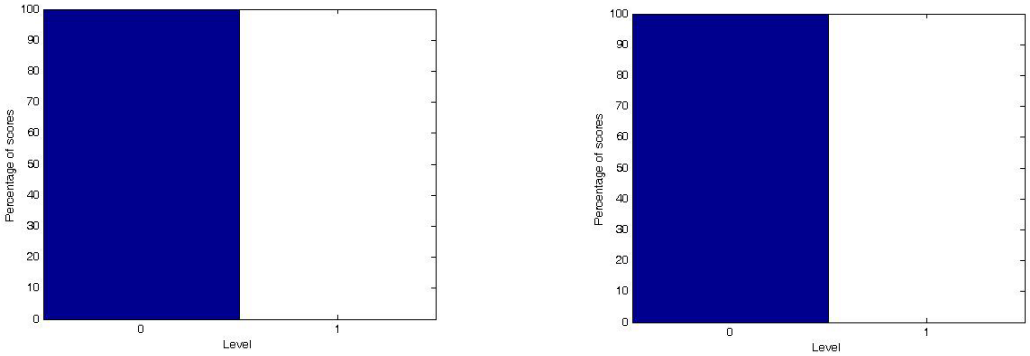


(c) Six-level pretest and posttest scores (p = .029)

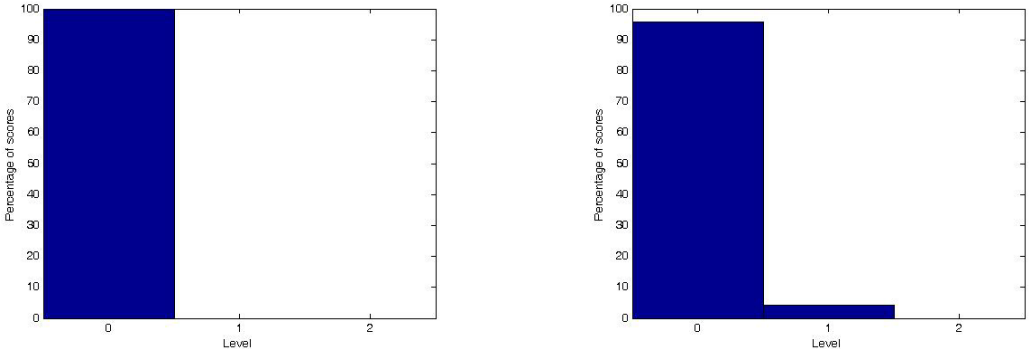


(d) Eleven-level pretest and posttest scores (p = .030)

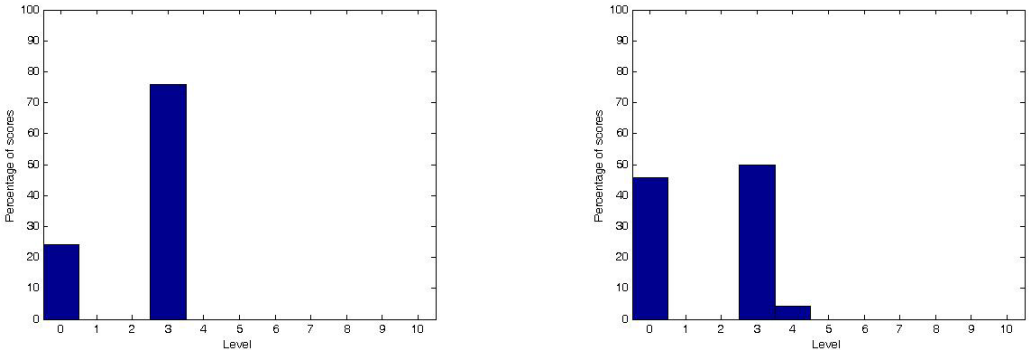
Fig. 8. Histograms for the target word "reject"



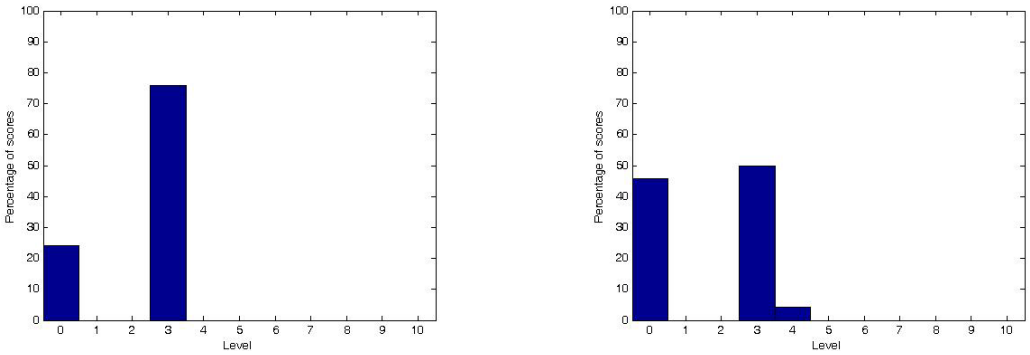
(a) Binary pretest and posttest scores (all zero scores, no comparison possible)



(b) Three-level pretest and posttest scores (p = .260)

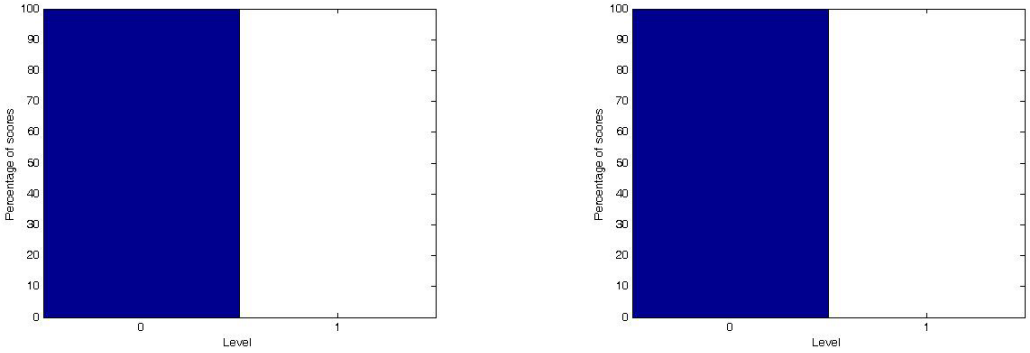


(c) Six-level pretest and posttest scores (p = .093)

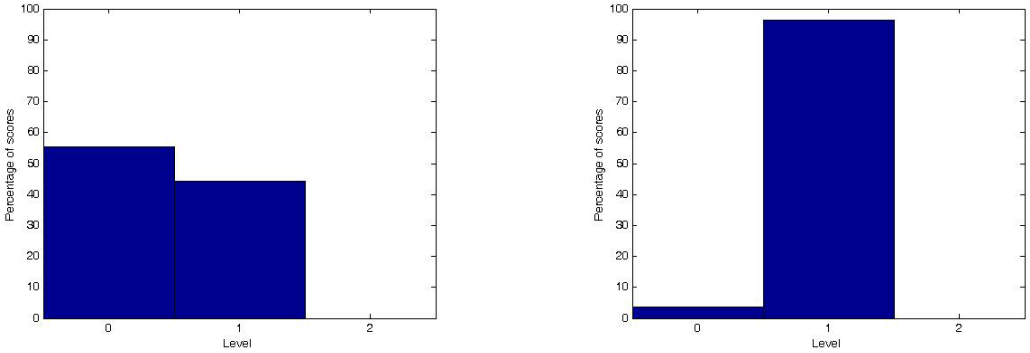


(d) Eleven-level pretest and posttest scores (p = .158)

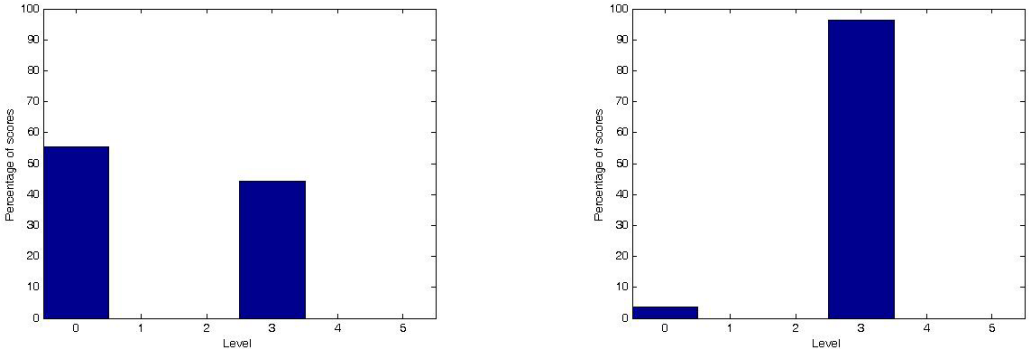
Fig. 9. Histograms for the target word "specify."



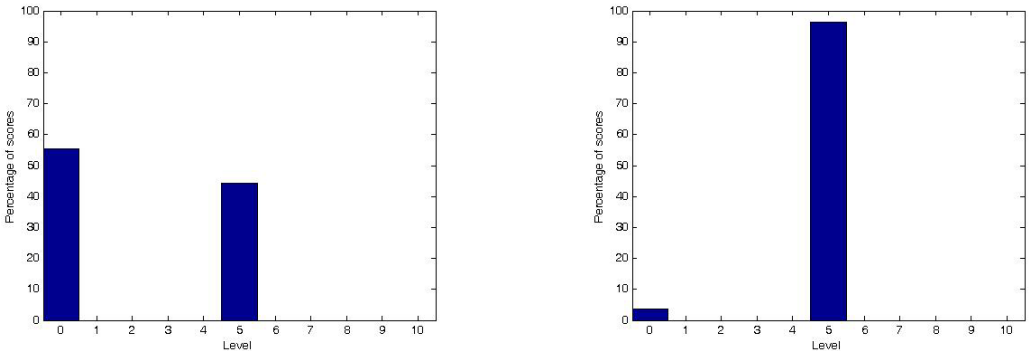
(a) Binary pretest and posttest scores (all zero scores, no comparison possible)



(b) Three-level pretest and posttest scores ($p \sim 0$)



(c). Six-level pretest and posttest scores ($p \sim 0$)



(d) Eleven-level pretest and posttest scores ($p \sim 0$)

Fig. 10. Histograms for a hypothetical target word

There is a further refinement of the scores at six- and eleven-level precision, where new PWK levels appear. The addition of the PWK scores is enough to bring the posttest improvement up to significance at the three-, six-, and eleven-level precisions.

The "all-zero problem." In addition to causing false positives and false negatives, low scoring precision made it impossible to carry out the *U* test for some target words. This problem occurred when subject target word knowledge was low in the pretest and the posttest, resulting in 0 scores for all translations in both tests. Table 7 lists the nine target words for which all-zero scoring occurred. The columns are arranged as in Tables 2 and 3. All 0 scoring is indicated by boldfaced, underlined text.

All-zero scoring did not always reflect rating precision. For two of the target words ("criteria" and "objective"), it occurred at all four levels of precision. This indicates that these words were difficult for the subjects. Many translations were given on the exams (8 for "criteria" and 21 for "objective"), all inaccurate.[4] On the other hand, for the other seven words in Table 7, the problem of all-zero scoring is related to low precision. For example, Figure 9 shows histograms for the word "specify." Figure 9(d) shows that the pretest and posttest scores are all between levels 0 and 5 – that is, in the lower half of the score range, between 0 and 2.5 points. As a result, with binary rating, the pretest and posttest scores are all 0. As the precision increases, nonzero scores appear in the histograms, and the all-zero problem disappears.

From Table 7, it is clear that none of the words for which the all-zero scoring problem arises underwent significant change in the posttest. However, such a situation (all-zero scoring and significant improvement) could potentially arise. To illustrate this point, Figure 10 shows histograms for a hypothetical target word. As shown in Figure 10(d), all of the translations for this imaginary word (both in the pretest and in the posttest) occur at either level 0 or level 5, corresponding to scores of 0 and 2.5. However, at the posttest there is a large change in the percentage of respondents who earn credit. During the pretest, fewer than half of the responses earn 2.5 points, whereas during the posttest all but one of them does, leading to a significant improvement. However, because the improvement takes place entirely within the lower half of the scoring range, all responses are rated 0 in the binary scoring system.

Summary of U test results. Figure 11 summarizes the relationship between scoring precision and *U* test results. The horizontal axis represents scoring precision and the vertical axis represents the number of target words. A total of 90 words were tested, and the subjects showed significant improvement in knowledge of 11 of them. Binary and three-level scoring failed to detect improvement in 4 of those words. In addition, binary scoring led to three false positives and three-level scoring led to two. However, no such errors occurred with six-level rating precision.

The advantages of multiple level scoring in assessment of PWK. Using the results obtained from the original, eleven-level scores as a standard, we have shown that binary and three-level scoring lead to a substantial number of errors. These include both false positives and false negatives. Furthermore, for several of the target words, binary and three-level scoring led to a "zeroing-out" of all the pretest and posttest data, making statistical comparison impossible. We conclude that the answer to research question 2 is "yes:" by recognizing more PWK levels, we obtain a clearer picture of the changes in the subjects' word knowledge.

We have argued for greater precision than has previously been used in the rating of L1 translations for assessment of PWK. To make our point, we did statistical analysis of pretest and posttest translation data under four different rating conditions: binary, three-level, six-level, and eleven-level

Table 7. Target words for which all of the pretest and posttest scores were 0

Target word	Pret avg. (N1)	Post avg. (N2)	2 lev	3 lev	5 lev	10 lev
			p	*p*	*p*	*p*
approach	0.00 (33)	0.02 (24)	**All 0**	**All 0**	.256	.256
coordinate	0.05 (31)	0.13 (32)	**All 0**	.340	.576	.576
criteria (*)	0.00 (28)	0.00 (28)	**All 0**	**All 0**	**All 0**	**All 0**
derive	0.09 (27)	0.32 (30)	**All 0**	.209	.169	.169
objective (*)	0.00 (32)	0.00 (23)	**All 0**	**All 0**	**All 0**	**All 0**
specify	1.14 (33)	0.83 (24)	**All 0**	.256	.093	.158
subsequently	0.09 (28)	0.00 (28)	**All 0**	**All 0**	.081	.081
technical	0.64 (32)	0.50 (32)	**All 0**	.527	.639	.651
underlying	0.00 (31)	0.16 (32)	**All 0**	.341	.086	.087
Total precision-related all-zero scoring			**7**	**2**	**0**	**0**

Note: (*) In these two cases, the problem of all-zero scoring is caused by word difficulty.

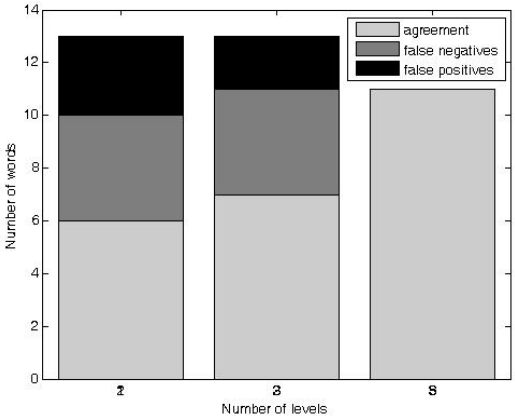


Fig. 11. Summary of the relationship between scoring precision and U Test results

scoring. Six- and eleven-level scoring led to different (and presumably more accurate) results than binary or three-level scoring.

Our rating procedure offers three advantages over previous, non-translation based approaches to PWK assessment. First, we compare only like objects: an L1 translation is compared only to other L1 translations. (There is no comparing of qualitatively different responses, as in the vocabulary knowledge scale.) This enables greater comparability and of the translation responses. After we pool together all the translations for a given target word, we list them together, on the same rating sheet. Our raters can use their intuitions to double check their work and ensure that the assigned scores reflect the relative accuracy of the translations.

A second advantage of our procedure is that it enables us to apply the same standard to the pretest and posttest. The ratings are done at the same time, after the completion of the posttest. At that time, all of the translations for a given target word are pooled together and presented to the raters, who are unaware of where they occurred (pretest and/or posttest). In our study, many of the pretest translations showed up again on the posttest. Our procedure ensured that a translation received the same value, whether it was used in the pretest or the posttest. There was no need to develop an assessment standard *a priori*, before the pretest, as in multiple choice testing.

A third advantage is that our approach is based on the qualities of the target words, not on assumptions about the vocabulary acquisition process that could vary in subtle ways from subject to subject.

In the future, new methods will be needed for carrying out multi-level translation rating in a more systematic and scientific manner. An interesting approach is reported in a study of L1 vocabulary acquisition among Dutch schoolchildren by Fukkink, Blok, and De Glopper (2001). The subjects in this study provided L1 definitions of the target words. The definitions were rated with the help of worksheets derived from dictionary information on the target words, including a list of semantic attributes. The scores were calculated using a formula that took into account the number of correct attributes exhibited by the definition, the number of false attributes, and the contextualization of the definition (the definition's dependence on the reading passages in which the students encountered the target word). There is a need for development of similarly innovative strategies in the rating of L1 translations.

REFERENCES

Anderson, R. & Freebody, P. (1983). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior*, *15*(3), 19-39.

Beglar, D. (2000). Estimating vocabulary size. *JALT Testing & Evaluation SIG Newletter*, *4(*1). 2-3.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*(2), 131-162.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Day, R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, *7*, 541-549.

Ferris, D. (1988). *Reading and second language vocabulary acquisition*. Unpublished manuscript, Department of Linguistics, University of Southern California.

Fukkink, R., Blok, H., & de Glopper, K. (2001). Deriving word meanings from context: A multicomponential skill. *Language Learning*, *51*(3), 477-496.

Gass, S. (1983). Language transfer and universal grammatical relations. In S. Gass, & L., Selinker (Eds.), *Language Transfer in Language Learning, Rowley* (pp. 69-82), Rowley, MA: Newbury House.

Glendinning, E., & Glendinning, N. (1995). *Oxford English for electrical and mechanical engineerin*g. Oxford University Press.

Glendinning, E., & McEwing, J. (2003) *Basic English for computing*. Oxford University Press.

Hermann, F. (2003). Differential effects of reading and memorization of paired associates on vocabulary acquisition in adult learners of English as a second language. *TESL-EJ (Teaching English as a Second or Foreign Language)*, *17*(1), 1-15.

Horst, M. (2000). Text encounters of the frequent kind: Learning L2 vocabulary through reading (Unpublished doctoral dissertation. Retrieved September 15, 2009, from http://www.lextutor.ca/text_encounters/

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

Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading, *Canadian Modern Language Review*, *56*(2), 308-328.

Hunt, A., & Beglar, D. (2005) A framework for developing EFL reading vocabulary. *Reading in a Foreign Language*, *17*(1), 23-59.

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, *21*(1), 47-77.

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

Kroll, J., & Curley, J. (1988). Lexical memory in novice bilinguals: The role of concepts in retrieving second language words. In M. Gruneberg, P. Morris, & R. Sykers (Eds.), *Practical aspects of memory: Current research and issues* (pp. 389-395). London: John Wiley & Sons.

Kroll, J., & de Groot, A. (1997). Lexical and conceptual memory in the bilingual: Mapping form to meaning in two languages. In A. de Groot, & J. Kroll (Eds.), *Tutorials in Bilingualism* (pp. 169-199). Mahwah, NJ: Lawrence Erlbaum.

Kroll, J., & Stewart, E. (1990). Concept mediation in bilingual translation. Paper presented at the 31st Annual Meeting of the Psychonomic Society, New Orleans, LA.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, *59*(4), 567-587.

Lockett, J., & Shore, W. (2003). A narwhal is an animal: Partial word knowledge biases adults' decisions. *Journal of Psycholinguistic Research*, *32*(4), 477-496.

Luppescu, S., & Day, R. R. (1993). Reading, dictionaries, and vocabulary learning. *Language Learning*, *43*, 263-287.

Meara, P., & Jones, G. (1990). *Eurocentres vocabulary size test* (version E1.1/K 10, MSDOS), Zurich: Eurocentres Learning Service.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly. 20*(2), 233-253.

Nagy, W., Anderson, R., & Herman, P. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, *24*, 237-270.

Nation, P. (1990). *Teaching and learning vocabulary.* New York, NY: Newbury House Publishers.

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

Reider, A. (2002). Implicit and explicit learning in incidental vocabulary acquisition. *VIEWS*, *12*(2), 24-39.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*, 589-619.

Saragi, T., Nation, P., & Meister, G. (1978). Vocabulary learning and reading. *System*, *6*, 72-80.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Swanborn, M., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261-85.

Swanborn, M., & de Glopper, K. (2002). Impact of reading purpose on incidental word learning from context. *Language Learning*, *52*(1), 95-117.

Waring, R. (2002). Scales of vocabulary knowledge in second language vocabulary assessment. paper published in Kiyo, The occasional papers of Notre Dame Seishin University.

Waring, P., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader?. *Reading in a Foreign Language*, *15*(2), 130-163.

Watts, J. (2002). The effect of multiple texts on vocabulary acquisition. Retrieved September 15, 2009, from http://www.tesolanz.org.nz./

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46-63.

Wesche, M., & Paribakht, T. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review*, *53*, 13-40.

Ulanoff, S., & Pucci, S. (1999). Learning words from books: The effects of read aloud on second language vocabulary acquisition. *Bilingual Research Journal*, *23*(4), 409-422.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, *57*(4), 541-572.

Received: Feb. 19, 2009 Revised: Aug. 31, 2009
Accepted: Nov. 07, 2009

Appendix 1. Word List for the Vocabulary Test

Number	Word	Sublist	Number	Word	Sublist	Number	Word	Sublist	Number	Word	Sublist
1	constant	3	24	output	4	47	interpret (*)	1	70	underlying (*)	6
2	retain	4	25	contrast	4	48	corresponding (*)	3	71	procedure	1
3	adequate	4	26	option	4	49	internal (*)	4	72	incident	6
4	investigate	4	27	derive (*)	1	50	trend (*)	5	73	adjust	5
5	access	4	28	administration (*)	2	51	focus	2	74	via	8
6	feature	2	29	overall (*)	4	52	react	3	75	sequence	3
7	approach (*)	1	30	contradict (*)	8	53	accurate	6	76	function	1
8	relevant (*)	2	31	resource	2	54	plus	8	77	constitute (*)	1
9	specify (*)	3	32	impact	2	55	transmit	7	78	criteria (*)	3
10	fundamental (*)	5	33	previous	2	56	contact	5	79	subsequently (*)	4
11	ignore	6	34	construct	2	57	concept (*)	1	80	contrary (*)	7
12	12. strategy	2	35	35. image	5	58	58. alternative (*)	3	81	81. emphasis	3
13	13. reverse	7	36	36. available	1	59	59. obviously (*)	4	82	82. eventually	8
14	14. insert	7	37	37. estimate (*)	1	60	60. exceed (*)	6	83	83. predict	4
15	15. technical	3	38	38. purchase (*)	2	61	61. ensure	3	84	84. release	7
16	16. period	1	39	39. implement (*)	4	62	62. aspect	2	85	85. expand	5
17	17. significant (*)	1	40	40. reject (*)	5	63	63. create	1	86	86. maximize	3
18	18. survey (*)	2	41	41. benefit	1	64	64. attach	6	87	87. conclude (*)	2
19	19. sufficient (*)	3	42	42. detect	8	65	65. transfer	2	88	88. imply (*)	3
20	$20. objective (*)	5	43	43. method	1	66	66. principle	1	89	89. sum (*)	4
21	21. indicate	1	44	44. similar	1	67	67. evident (*)	1	90	90. eliminate (*)	7
22	22. enable	5	45	45. consist	1	68	68. coordinate (*)	3			
23	23. flexible	6	46	46. remove	3	69	69. substitute (*)	5			

Note: (*) control word.