An Exploratory Study on the Assessment of Semantic Knowledge of English Polysemy

Li-Szu Agnes Huang
National Kaohsiung First University of Science and Technology

Abstract
Using a combination of quantitative and qualitative paradigms, this study explored the possibility to measure Chinese EFL learners' semantic knowledge of English polysemy. Participants were 49 English majors from a Taiwanese university. The quantitative paradigm involved the test formats of translation and multiple-choice questions that measured participants’ comprehension of unfamiliar word senses by reading disambiguation cues. Participants were also required to self-rate their confidence in each response. The qualitative paradigm comprised an interview that elicited participants’ comments/perspectives on their knowledge gains from cue processing. The results of quantitative and qualitative data analyses revealed a considerable consistency in cueing effect. The translation test was observed to be more difficult while the multiple-choice test more sensitive to knowledge gains. Additionally, the design of cues was found to impact the validity of the translation test. It was suggested that to well detect EFL learners’ knowledge of English polysemy, the multiple-choice distractors should be conceptually and semantically close to the target word. The unique procedures of test design, operationalization, and administration were discussed to highlight the usefulness and constraints of the above self-designed measures. Also addressed were difficulties to develop tests that effectively assess learners’ semantic knowledge of English polysemy.

Keywords: language testing, semantic knowledge, English polysemy

INTRODUCTION
Knowing the meaning range of a word is generally recognized as an essential component of overall lexical proficiency. For second language (L2) learners, learning a word in L2 should involve acquiring the different meanings or senses associated with a word and grasping the concept underlying its related uses. In this regard, polysemy, a single linguistic form with multiple related meanings, constitutes an important dimension of depth of L2 word knowledge (Carter, 1998; Nation, 2001). Knowledge of polysemy also serves as a useful index for predicting EFL (English as a foreign language) learners’ performance on academic reading (Qian, 2002). Unfortunately, among the various measures of word knowledge, very few target at L2 polysemy.
The scarcity of such measures may stem from the difficulty in accurately assessing L2 learners’ receptive and productive knowledge of polysemous words. Unlike individual lexical items whose meanings are clearly distinguished from each other, the different senses of a polysemous word do not always have a clear-cut boundary in their meanings. Such a characteristic makes it difficult to design a context that is transparent enough to cue only one sense. Accordingly, great care should be taken in the design of test items and contextual information for cueing specific senses. In addition, the assessment of productive knowledge is challenging because learners tend to avoid the use of unfamiliar senses in language production. So far, hardly any measures are designed for this special purpose.

Due to the difficulties in finding appropriate measures, this paper explored the ways to assess Chinese EFL learners’ semantic knowledge of English polysemy. Specifically, the method presented cues for disambiguating unfamiliar senses first, and then examined learners’ gains in knowledge of these senses after reading the cues. Both quantitative and qualitative measures were applied to assessing the gains in receptive knowledge of target senses. It was hoped that the results of this study would serve as a stepping-stone for future research on the assessment of knowledge of L2 polysemy.

**LITERATURE REVIEW**

**Avenues to Detect Knowledge of L2 Words**

Concerning the measures of word knowledge in the L2, a distinction is often made between breadth and depth of vocabulary knowledge. While the vocabulary breadth tests estimate vocabulary size, the depth tests focus on the quality of vocabulary knowledge. Because the assessment of knowledge of L2 polysemy is a type of depth tests, the following literature review focuses on assessment tools for measuring the depth of vocabulary knowledge.

The depth tests often vary widely in design due to different aspects of vocabulary knowledge measured and different rationales for the construct of word knowledge. A well-known test for assessing general English word knowledge for L2 learners is the TOEFL test. The vocabulary section of the test adopts a context-dependent format to measure knowledge of the semantic features of a target word embedded in a reading passage. It also assesses test-takers’ strategies of guessing word meanings from context (Read & Chapelle, 2001). Among the wide range of words covered in TOEFL, words with multiple meanings/senses have also been tested. However, Schmitt (1999) questioned the construct validity of its vocabulary items by interviewing thirty L2 learners of English immediately after they took the TOEFL test. He found that the vocabulary items did not adequately reflect the participants’ knowledge of a word’s associations, grammatical properties,
collocations, or meaning range. Concerning the meaning range of a polysemous word, learners tended to know the most common sense productively, but only one or two other senses receptively even though they answered the test items correctly. In this regard, the TOEFL test does not seem an ideal tool for the measurement of semantic knowledge of English polysemy.

Another common measure for examining the semantic knowledge of L2 words is the test designed by Read (1993). Having a word association format, the test aims to assess L2 learners’ knowledge of association among semantically related words. The test format comprises a stimulus word and eight words that consist of distractors and associates of either a paradigmatic, syntagmatic, or analytic type. Adopting such a format to Dutch learners of French, Greidanus and Nienhuis (2001) found that stimulus words with higher frequency tended to elicit better knowledge of the tested aspects. However, it is suspected that learners may select different associates if they attend to different meaning dimensions of a polysemous stimulus word. Accordingly, a highly polysemous word may not be an ideal stimulus item for the test. Such a format, unfortunately, does not seem effective in measuring knowledge of L2 polysemy.

A possible option is to adopt the format of the Vocabulary Knowledge Scale, which combines self-report on receptive knowledge and performance items on productive knowledge of particular English words (Wesche & Paribakht, 1996). It has the advantage of measuring both receptive and productive knowledge at one time. Nonetheless, the designers clearly stated that the scale aimed to track the vocabulary development of ESL (English as a second language) learners in a university setting rather than estimate specific aspects of word knowledge. Other commonly used vocabulary depth tests, as reviewed in Read and Chapelle (2001), focused on lexical proficiency in written composition or in oral interaction. None of them targets at knowledge of polysemy. As no existing measures of L2 word knowledge can meet the current research need, is it possible, then, to adopt the test methods for measuring first language (L1) users’ knowledge of polysemy? The following section reviews some common methods used in an L1 context.

**Test Methods for Assessing Knowledge of L1 Polysemy**

A common and easy way to induce L1 language users’ knowledge of polysemy is to ask them to produce as many senses of a polysemous word as they can. For example, participants in Raukko (1999) were required to write down sentences exemplifying all the possible uses of the polysemous word *get*. Then they had to comment on the centrality of senses and the relatedness among the senses they had provided. The purpose was to build up native speakers’ intersubjective perception of the polysemous characteristics of *get*. However, such a method may not be suitable for L2 learners because they may produce only a few senses due to their limited exposure to a polysemous word in different contexts. They
are also likely to categorize similar senses together (Huang, 2003) and accordingly fail to provide a satisfactory number of different senses.

Another common test format is a sorting task that aims to discover L1 speakers’ perception of sense relations. The participants are asked to sort out a given set of senses of a polysemous word either freely or into per-determined categories (Baker, 1999; Gibbs & Matlock, 1997). The main purposes are to see what principles test-takers use to classify senses and what relations they perceive among senses. For the latter purpose, an alternative instrument, a semantic relatedness scale, is also often used. A 5- or 7-point scale is provided for respondents to rate the degree of similarity between any pair of senses (Durkin & Manning, 1989). The above two methods, though worked well in eliciting subjectively perceived sense relations, may not be suitable for L2 learners. One reason is that unlike L1 speakers who readily see the connections among word senses due to their extensive exposure to a polysemous word in a variety of contexts, L2 learners may lack the sensitivity to these sense relations. In addition, as learners tend to understand senses using L1 translation equivalents, their judgments on sense relatedness are likely to differ according to various L1 backgrounds. Particularly for learners whose L1 has a distinct orthographic system from the L2, they may fail to notice the connections among senses if the translation equivalents of these senses turn out to be discrete (Huang, 2003). Using the perception of sense relatedness to infer learners’ semantic knowledge of L2 polysemy may create a skewed picture.

The aforementioned test methods for L1 speakers have provided valuable data for researchers to discover the motivation for sense extensions and to depict the behavior of individual polysemous words. However, if these tests are to be applied to L2 learners, revisions may be needed because test designers have to consider the following factors that are characteristic of measures on L2 polysemy.

**Features of the Measurement on L2 Polysemy**

The first factor concerns the design of context for a target sense. As polysemy usually arises from small extensions of the core sense from which other meanings are computed in a variety of contexts (Clark, 1993), context is crucial in determining the intended meaning of a polysemous word. To put it differently, the words surrounding a polysemous word create a context that “selects” a particular meaning dimension of the word. Hence, context is especially important in the assessment of receptive knowledge of L2 polysemy. For example, when investigating Dutch ESL learners’ knowledge of figurative senses of English polysemous words, Verspoor and Lowie (2003) used a translation test in which each target polysemous word was embedded in a sentential context. Despite the essential role of context, the amount and the degree of transparency of contextual information are subject to different assessment purposes.
The second factor is learners’ L1 backgrounds. Some languages such as English and Spanish have a high density of cognates that share a certain number of polysemous readings. Using cognates as test items in a lexical decision task allows researchers to test what sense is activated first and in which language mode (Beauvillain & Grainger, 1987; De Groot, Delmarr, & Lupker, 2000). However, designing a similar test for learners whose L1 differs widely from the target language may be challenging. Although some researchers tried the lexical decision task on Chinese learners of English, the primes or targets were usually single Chinese characters (Chen, 1990; Chen & Ng, 1989; Keatley, Spinks, & Gelder, 1994). To express different English senses in Chinese, it is necessary to present a Chinese lexical item that usually consists of more than one character. In some cases, test items that describe specific senses have to be presented in sentences as primes or targets. Under such circumstances, it would be difficult to control the degree of complexity of each prime and target. Consequently, a lexical decision task format is almost impossible to be used for measuring the processing patterns of Chinese learners of English.

As an exploratory study, this paper reported experimental measures to assess Chinese EFL learners’ gains in receptive knowledge of unfamiliar English verb senses after reading disambiguation cues. A translation and a multiple-choice test were used to generate quantitative data and an interview was conducted to obtain qualitative data. The study aimed to see if such a testing method could faithfully capture learners’ semantic knowledge of English polysemy.

METHODOLOGY

Participants

Participants were 49 English-major sophomores at a university in Taiwan. Eight of them participated in an interview session. They were all Chinese speakers who had learned English as a foreign language. By the time they began their university studies, they had had at least six years of English instruction in high school. Although students were not given a formal measure of their vocabulary knowledge in English, their level of vocabulary knowledge was estimated as high-intermediate based on two assumptions. First, as English
majors, their scores on the English subject section of the Joint Entrance Exams to Colleges needed to be at a certain level before they could be admitted to the university, one of the most prestigious universities in Taiwan that typically selects students of a high level of English proficiency. Second, based on the course materials they were able to read at the time and the sound judgment of their classroom instructors, these students had acquired the amount of vocabulary that high-intermediate EFL learners should know. Therefore, they could be categorized as high-intermediate learners of English in terms of their vocabulary knowledge and overall English proficiency.

**Test Format**

Prior to the experiment, a preliminary test had determined a list of 18 unfamiliar senses of English polysemous verbs for the target group of learners. The test instruments for the current experiment consisted of cues for disambiguating unfamiliar English senses and quantitative and qualitative measures that assessed participants’ knowledge gains after reading the cues. First, four cue conditions were created for each test item: elaborated context, semantic frames, meaning chains, and the control condition with no cues provided. The elaborated context cue consisted of two sentences that provided rich contextual information for understanding the target sense. The semantic frame cue was composed of an English sentence with a cueing verb for activating the underlying concept of the target sense. The meaning chain cue comprised two or three sentences with one representing the core sense of the target polysemous word and the others a sense connecting the core sense and the target sense. These three types of cues were hypothesized to help disambiguate the intended sense of a polysemous word, but their effects were yet to be confirmed.

The quantitative measures included a translation and a multiple-choice test format. A translation test was used because of its usefulness as an elicitation technique for assessing second language learning (Haastrup & Henriksen, 1998; Read, 2000). The multiple-choice format was selected because it is able to control the required responses and is easy to administer, score, and analyze (Genesee & Upshur, 1996). In the current study, the translation test not only measured directly learners’ understanding of the meaning of the target sense but also provided a chance for learners to produce an answer that is not included in the multiple-choice questions. The multiple-choice format, on the other hand, tends to be more sensitive to detect the amount of knowledge of target items (Nation, 2001). To counterbalance the possibility of guessing in the two tests and to estimate the demonstrated knowledge precisely, a 5-point Likert scale was provided additionally for learners to self-report their confidence in each response generated. The confidence ratings then served as an index of participants’ self-assessment of knowledge gains.

Each test item was presented in a sentential context with limited clues to its intended meaning. The translation test had such a sentence with the target sense underlined and a
An Exploratory Study

信心度量表。多重选择题包含一个带有测试词义的句子，一个包含四个选项的多重选择题，以及一个信心度量表（请参阅附录以获取测试格式）。四个选项包括一个正确答案，目标多义词的核心意义，一个替代词义，以及一个无关意义。线索和两个测试被结合并打印在8×3.5英寸的纸张上形成一个小册子。每个小册子包含18个测试项目，每个项目在第一张纸上有特定类型的线索，在第二张纸上有翻译测试，在第三张纸上有多项选择题。

质性测量是一个访谈，用于引起参与者对知识获得的评论。此研究未使用口头报告，一种理解心理过程的常见方法，因为试验中参与者无法大声说出他们的想法而进行测试。他们的陈述并未提供有关他们理解过程的有用数据。因此选择了访谈。访谈问题关注于参与者如何使用线索来找出目标词义以及如何回应测试问题。这样的口头报告可以显示受访者的对测试项目的知识程度，以及提供宝贵的证据来验证定量测试的结构验证（Bachman，1990；Schmitt，1999）。

测试程序和数据分析

所有参与者，除了八个学生用于访谈外，同时在他们的常规教室中进行测试。十名学生被随机分配到三个条件中的每一个，而另外十一名则被分配到控制条件。完成一组练习后，每个参与者独自处理包含18个测试项目的册子。学生们被不断地提醒不要翻转测试纸张。整个测试大约需要45分钟。

另外八名学生在另一个教室中单独进行测试。两人为一组被随机分配到四个条件。在访谈中，每个参与者先做练习，然后处理测试项目。完成每个项目后，参与者被停止并要求反思他们当时在做什么。由于时间限制，三个条件中的每个参与者被访谈了总数的一半项目。相比之下，两个控制条件的参与者被访谈了整个项目集。每个访谈持续约30到40分钟。访谈后，学生完成了剩余的项目。所有访谈都进行了录音。

研究者将参与者的定量和定性测量的反应视为研究数据。首先，计算每个条件中两个测试的正确比例均值和信心度量表的均值。对于167
confidence ratings, the mean scores for correct and incorrect answers were obtained separately. The number circled in the scale was assigned as the score for a particular item regardless of the accuracy of answers. Then, paired samples t-tests were carried out for each cue condition to compare the results generated in the two tests. One-way ANOVA's were conducted to detect the effects of cues on participants’ accuracy rate and confidence rating scores for each test.

To examine if the performance in the two tests was consistent, conditional probability of the two tests was obtained. That is, four types of statistics were generated for each cue condition: the percentages of participants who answered both tests correctly, who answered both tests incorrectly, who answered correctly in translation but incorrectly in multiple-choice, and who answered correctly in multiple-choice but incorrectly in translation. As for the qualitative data, the recorded interviews were transcribed and analyzed to examine the degree of consistency between the quantitative and qualitative results.

RESULTS

Quantitative Results

The results of the proportion of items answered correctly in the two tests are tabulated in Table 1. The first panel shows that for both tests, Cue A elicited the highest means for both tests whereas Cue C the lowest. Cue B generated a lower mean in translation but a slightly higher mean in the multiple-choice test than the control condition. The comparisons of cue effects in the lower panel, however, identify significant supremacy only for Cue A against Cue D. That is, Cue A had the strongest disambiguation effect as measured in the multiple-choice test.

It was surprising that students receiving no cues could achieve more than 50% of accuracy in translation. Because the items were supposed to be unfamiliar to students, it is suspected that students managed to guess the answer correctly by consulting with the sentence in which the target sense was embedded. Because this study did not include English vocabulary knowledge or reading ability as a covariate, it is difficult to judge if participants’ level of reading proficiency was also responsible for the results reported here. For future improvement, the research design would need to take into consideration learners’ vocabulary knowledge or reading proficiency in English.

In terms of performance in the two tests, the participants generated an overall higher mean score in the multiple-choice test than in the translation test. This difference suggested that the multiple-choice test was easier and also better in detecting knowledge gains. The significant results found in Conditions A and B revealed that participants in these two conditions improved their performance impressively in the multiple-choice test. Because
Cue A was significantly better than the no-cue condition for this test, students’ better performance may be partly attributed to the strong cue effect. In comparison, the higher mean for Cue B in the multiple-choice test was more likely to result from the ease of this test rather than cue effects. On the other hand, Cue C, which elicited lower means, though not significantly lower, than the control condition, did not seem to promote students’ performance in either test. In a word, different natures of tests and cue effects both contributed to the test results.

Concerning confidence ratings for correct and incorrect answers, the overall $F$ tests on cue effects did not reach a significant level of .05 for either correct answers ($F(3,45) = 2.18, p = .145$ for translation and $F(3,45) = 1.73, p = .029$ for multiple-choice) or incorrect answers ($F(3,45) = 1.31, p = .283$ for translation and $F(3,45) = 0.73, p = .542$ for multiple-choice). Accordingly, the comparisons among cue conditions were not pursued further. In other words, cue types did not significantly influence participants’ confidence ratings in either correct or incorrect answers. So Table 2 only presents the descriptive statistics and $t$-test results.

As shown in the first panel, the mean ratings for correct answers in the multiple-choice test were higher than those in the translation test in all conditions except Condition B. Students receiving Cue B reported a slightly higher confidence score in their translation responses than in their answers to multiple-choice questions. It is suspected that with a cueing verb serving as a synonym or near-synonym for the target sense, students may feel more confident in using the translation equivalent of the cueing verb as a response. However, the mean difference did not reach a significant level in any of the conditions. In a word, the participants in all conditions actually had similar levels of confidence when their answers turned out to be correct.

### Table 1. Proportion of Correct Items in Two Tests

<table>
<thead>
<tr>
<th>Cue condition</th>
<th>Translation</th>
<th>Multiple-choice</th>
<th>Comparison of two tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>Cue A (N=12)</td>
<td>0.59</td>
<td>0.09</td>
<td>0.72</td>
</tr>
<tr>
<td>Cue B (N=12)</td>
<td>0.47</td>
<td>0.16</td>
<td>0.60</td>
</tr>
<tr>
<td>Cue C (N=12)</td>
<td>0.49</td>
<td>0.13</td>
<td>0.53</td>
</tr>
<tr>
<td>Control (N=13)</td>
<td>0.54</td>
<td>0.14</td>
<td>0.59</td>
</tr>
<tr>
<td>Overall</td>
<td>0.52</td>
<td>0.14</td>
<td>0.61</td>
</tr>
</tbody>
</table>

**Comparison of cue conditions**

<table>
<thead>
<tr>
<th></th>
<th>Translation</th>
<th>Multiple-choice</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A vs. D</td>
<td>0.75</td>
<td>.391</td>
<td>5.17</td>
<td>.028*</td>
</tr>
<tr>
<td>B vs. D</td>
<td>1.81</td>
<td>.185</td>
<td>0.11</td>
<td>.731</td>
</tr>
<tr>
<td>C vs. D</td>
<td>0.88</td>
<td>.253</td>
<td>0.67</td>
<td>.422</td>
</tr>
<tr>
<td>Overall</td>
<td>1.89</td>
<td>.145</td>
<td>3.30</td>
<td>.029*</td>
</tr>
</tbody>
</table>

**Note.** Cue A: elaborated context. Cue B: semantic frames. Cue C: meaning chains.
As for incorrect answers, the lower panel of the table shows no significant difference in the means for the two tests in most conditions except in the control condition. That is, students receiving cues rated their confidence level equally for both tests when their answers turned out to be incorrect. By contrast, students receiving no cues tended to report more confidence in taking the multiple-choice test even though they scored incorrectly. It is possible that multiple-choice questions provided additional clues and rendered students bolder to guess. Generally speaking, a higher mean was found for correct answers than for incorrect ones across cue conditions and test formats. Such a phenomenon suggested that students tended to report a higher level of confidence when they actually answered correctly; they were less confident in the answers that eventually proved to be wrong.

<table>
<thead>
<tr>
<th>Table 2. Confidence Ratings on Correct and Incorrect Answers on Two Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correct Answers</strong></td>
</tr>
<tr>
<td><strong>Cue condition</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Cue A (N=12)</td>
</tr>
<tr>
<td>Cue B (N=12)</td>
</tr>
<tr>
<td>Cue C (N=12)</td>
</tr>
<tr>
<td>Control (N=13)</td>
</tr>
<tr>
<td>Overall</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Incorrect Answers</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cue condition</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Cue A (N=12)</td>
</tr>
<tr>
<td>Cue B (N=12)</td>
</tr>
<tr>
<td>Cue C (N=12)</td>
</tr>
<tr>
<td>Control (N=13)</td>
</tr>
<tr>
<td>Overall</td>
</tr>
</tbody>
</table>

*Note.  Cue A: elaborated context.  Cue B: semantic frames.  Cue C: meaning chains.*

To compare students’ performance in the two tests, the conditional probability of the two tests in each condition is plotted in Table 3. If two tests are consistent in measuring the same ability, high scores should fall in the cell of correctness for both tests and the cell of incorrectness for both tests. Most of the conditions did show such a pattern. Particularly in the control condition, the sum of the percentages in the two cells of correctness and incorrectness for both tests is much higher than that in the other two cells, suggesting a high degree of consistency in performance. By contrast, students in Condition C had less consistent performance as approximately 40% of them answered correctly in one test but incorrectly in the other. For conditions A and B, almost one-fourth of students scored wrongly in translation but accurately in the multiple-choice test. So in general, more students in these two conditions scored correctly in the multiple-choice test than in the translation test. Such results converged with those in Table 1: participants receiving Cues A
and B performed significantly better in the multiple-choice test than in the translation test.

### Table 3. The Conditional Probability of Two Tests

<table>
<thead>
<tr>
<th>Condition</th>
<th>Translation</th>
<th>Multiple-choice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
</tr>
<tr>
<td><strong>A</strong></td>
<td>47.7%</td>
<td>11.1%</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>35.6%</td>
<td>11.1%</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>31.0%</td>
<td>18.1%</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>41.4%</td>
<td>12.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
<th>Translation</th>
<th>Multiple-choice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>24.1%</td>
<td>17.1%</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>22.2%</td>
<td>28.7%</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>16.7%</td>
<td>29.1%</td>
</tr>
</tbody>
</table>

**Note.** The number in each cell represents the percentage of participants.

### Qualitative Results

The qualitative data represented students’ self-report on their cue processing and test-taking strategies. Their accounts on cue processing revealed that participants receiving the elaborated context cue were able to use it at ease to understand the target sense. By contrast, those receiving the cues of semantic frames and meaning chains frequently reported difficulties in connecting the cues with the target sense. These comments were not surprising because contextual information is often the most natural and readily accessible resource for L2 learners to deal with unfamiliar lexical items (Nagy, 2001). Processing the other two types of cues actually placed more cognitive load on the participants; they had to process the cues deeply to grasp the underlying word concept or to discover the connections among senses. The cue of elaborated context, therefore, resulted in the strongest disambiguation effect whereas the other cues did not show prominent effects.

The way the two participants in the control condition dealt with unfamiliar senses was a little different. Without any disambiguation cues, they could only rely on their existing knowledge of the target polysemous verbs and the limited contextual information provided by the target sentence. The more competent participant of the two actually used other words in the sentence or the grammatical structure as clues to guess the target sense. Both participants used their L1 to extend from the core sense, but because their L1 elaborations varied, their translation responses differed more dramatically than the answers to the multiple-choice questions. In fact, these two interviewees reported that the four choices in the questions did help them to narrow down their wild guesses. Such a tendency explained why students had higher confidence ratings for the multiple-choice test even though their answers turned out to be wrong (also shown in Table 2).

Similarly, respondents receiving cues reported multiple-choice questions as helpful in correcting their misunderstanding. When their understanding of the target sense was challenged by a mismatch between their translation and the four choices in the question, they
tended to refine their original interpretation to match one of the choices. Also observed was
a greater amount of time spent on doing the translation test. Some participants even
provided deviant or inappropriate translation responses for certain items although their oral
report showed correct understanding of the items. The difficulty in taking the translation
test echoed the quantitative results of lower means for this test (see Table 1).

In addition to the influence from cue effects and natures of tests, the design of cues also
impacted the validity of the translation test. Particularly in the semantic frame condition,
better disambiguation effects were found in items that used a synonym as the cueing verb.
These items thus elicited better translation performance than those using a non-synonymous
cueing verb. For the other two types of cues, if the cue created ambiguity to the extent of
misleading students, the possibility for inaccurate translation responses obviously increased.
The interaction of these factors accounted for a higher degree of inconsistency found in the
three cue conditions than in the control condition (please refer to Table 3).

DISCUSSION AND CONCLUSION

In this study, the participants went through a process of reading a type of disambiguation
cue, taking a translation and a multiple-choice test, and self-rating the confidence level in
each answer. Those who participated in the interview session were requested to reflect on
the above procedure. Could the use of these methods provide a clear picture of learners’
gains in knowledge of unfamiliar word senses? Both quantitative and qualitative results
revealed a considerable consistency in cue effect. That is, the elaborated context cue created
the greatest ease in processing and the best performance across tests. The knowledge gains
from processing the semantic frame cue were better detected in the multiple-choice test,
whereas the meaning chain cue had an effect too weak to result in any knowledge gains.
Generally speaking, results measured by the two tests varied in different cue conditions
because of the complicated interactions among cue design, cue effect, and test natures.

Influences of Test Formats

The translation test, with limited clues to the meaning of the target senses, appeared to
be more difficult than the multiple-choice test, a format that is likely to provide a certain
extent of clues that facilitated guessing. It was observed in the interview session that
sometimes participants had difficulty finding an appropriate translation for the meaning they
had discovered; such a meaning was then found in the multiple-choice question. In other
cases, students did not know the meaning they had figured out was wrong until they realized
that none of the four choices matched their translation. Their understanding was somehow
verified or corrected in the multiple-choice test. This explained why participants had
inconsistent performance in the two tests. Especially for those receiving Cues A and B, the cue effects together with the clues from multiple-choice questions contributed to better performance in the multiple-choice test.

Indeed, the nature of different test formats is likely to influence test-takers’ performance (Shohamy, 1984). Comparing with a multiple-choice format that measures receptive knowledge, a translation test targets more at the productive knowledge and thus calls for precise understanding of the target word (Nation, 2001). In this study, a certain number of translation responses were marginally correct, suggesting that students may hold partial understanding of the target sense. Adding to the problem was learners’ level of L1 competence: some were unable to express their correct understanding of the target sense with an appropriate L1 equivalence. In addition, the validity of the translation test was challenged by the design of cues, especially the selection of the cueing verb in the semantic frame condition. All these observations implied that simply using a translation test may not be sufficiently effective in assessing L2 learners’ semantic knowledge of polysemy.

In comparison, the multiple-choice test format seemed better in detecting knowledge gains in this study. However, the test had its disadvantages in the design of distractors. For some items, students only needed to decide between the correct answer and another sense of the target word. Under such circumstances, the test became less powerful in discriminating the degrees of understanding learners had achieved. As proposed in Herman, Anderson, Pearson, and Nagy (1987), multiple-choice distractors that are conceptually and semantically close to the target word may work better in measuring the target dimension of knowledge. Therefore, when assessing L2 learners’ knowledge of polysemy, test designers need to adopt a higher level of difficulty for the distractors. This practice would help to reduce the possibility of guessing and to increase the power of the test.

Functions of the Qualitative Paradigm

The use of an interview to elicit participants’ report on cue processing and test-taking strategies functioned to triangulate the quantitative data. Although the number of interviewees was not large enough to form a pattern of responses, they did provide useful accounts about their comprehension process and reasons for responses. Their comments also highlighted possible flaws in cue design that had affected the disambiguation power of cues. For example, the participants perceived the senses used in the meaning chain cue as having weak or no relations with the target sense. The mediating sense in the cue, in some cases, even produced confusion that seriously impaired comprehension. Such an observation revealed that due to little help from the cues, students in this cue condition performed poorly.

In addition, the qualitative data allowed the researcher to disentangle the mystery of participants’ inconsistent performance in the two tests. The three types of cues,
hypothesized to help disambiguate unfamiliar word senses, became less powerful when participants failed to interpret the cues correctly. While the quantitative results only signaled the weakness of certain cue types, the qualitative data disclosed participants’ misinterpretations of some cues, a possible reason for the different results in different conditions.

However, the use of an interview had its drawbacks. It was observed that while doing the first several items, interviewees did not always know the function of the cues, just like the other participants who took the test on their own. Then, through interviewing questions, some seemed to “catch the point” and exhibited a better skill in connecting the cue with the target item for the rest of the test. Therefore, receiving interviews during the test-taking process may not resemble the actual comprehension process wherein no external guidance or interference was present.

Constraints of the Experimental Measures

The current testing methods successfully measured students’ knowledge gains after reading cues. However, there are some constraints of these experimental measures. The first constraint lies in the design of cues and test questions. To ensure the validity of the assessment instruments, the multiple-choice distractors should be equally attractive and complex; each cue under the same condition needs to be similar in its disambiguation power. Unfortunately, such an ideal situation was not successfully created despite of great care in cue design. For some items, ambiguity found in the cues caused misunderstanding of the cues. This problem is not easy to solve, though. As senses of a polysemous word do not always have a clear boundary in meaning, it is very difficult to avoid ambiguity in a test measuring the understanding of specific senses. It is neither easy to ensure all test-takers have the same level of knowledge of the words used in the cues or test questions.

The current study followed the rules of operationalization (Bachman & Palmer, 1997) and tried to simulate the authentic situation of applying disambiguation resources to understanding an unfamiliar sense. What may not be authentic was the limited context in which a target sense was embedded. Such a design was a compromise with the additional purpose of testing cue effects. In order to know if the cues helped, it was necessary to use a deliberately neutral context so that only restricted clues were provided for the meaning of the target items. The construct of the test thus focused on knowledge gains after reading the cue. Another constraint in operationalization was the scoring of translation responses, which was done by the researcher alone. Because a translation test often elicits a range of answers, it is better to have multiple raters to establish a good inter-rater reliability in scoring (Read, 2000).

In test administration, the impossibility to control the time for cue processing may partly account for the weak power of cues. For the success of the cues of semantic frames and meaning chains, a deep processing of the cues is critical. Unfortunately, some participants
simply read through the cues without thinking of using the cues to solve their puzzle of the target sense. It is suspected that if tested with a computer program that controlled the time for cue presentation, students may be more likely to process the cues deeply. The disambiguation power of cues may increase so as to induce significant knowledge gains. However, due to some technical problems and unavailability of such a computer program, paper-and-pencil tests were used instead. Future research could take such an issue into account.

Despite the above limitations, this study intended to act as one of the pioneer studies on L2 learners’ semantic knowledge of English polysemy. Researchers interested in this area are encouraged to develop tests that work effectively toward this end. This should be a sensible and meaningful mission given the enormous difficulty researchers may encounter in designing appropriate measures.

REFERENCES


Appendix: An Example of Test Items and Test Format

**Target verb: bring**

**Condition A** (elaborated context):
My conversation with Jane came to a surprising end. I asked her about her life as a young kid.

**Condition B** (semantic frame):
The gunman forced us into the room.

**Condition C** (meaning chain):
(a) Remember to bring me a book.
(b) What brings you here? What causes you to come here?

**Condition D** (control):
Nothing is provided.

**Translation task**
She could not bring herself to talk about her childhood.
Chinese translation __________
How are you confident that your answer is correct?
Not confident at all Very confident
1 2 3 4 5

**Multiple-choice task**
She could not bring herself to talk about her childhood. She ___ her childhood.
(1) couldn’t remember
(2) had nothing to say about
(3) didn’t keep anything from
(4) didn’t want to discuss
Answer __________
How are you confident that your answer is correct?
Not confident at all Very confident
1 2 3 4 5