Japanese EFL Learners’ Weaknesses and Strengths in Basic Vocabulary Knowledge: Focusing on Good and Poor Spellers

Sachiyo TAKANAMI
Toyo University

Abstract

The main purpose of this paper is to ascertain the findings from Takanami’s (2014, 2015) studies on the difficulties of spelling acquisition among Japanese EFL learners. A total of 34 Japanese university students participated in the present study. Ten test formats including six recall tasks and four recognition tasks, focusing on three fundamental abilities (i.e., decode, spell, and comprehend), were conducted for each target word to examine weaknesses and strengths in basic vocabulary knowledge among the participants. The target words were 33 high-frequency, already learned words. Test-takers’ estimated vocabulary size was tested beforehand. The results of this study confirmed that (a) the productive knowledge of the written form was the most serious deficiency in the basic knowledge of vocabulary acquisition among the learners, (b) 38% of the learners were in the category of poor spellers (S-), and (c) good spellers (S+) are always good decoders (D+) and good comprehenders (C+). This paper concludes that the various test formats succeeded in showing the learners’ weaknesses and strengths in basic vocabulary acquisition. Thus, specifically designed test formats, which assess learners’ basic knowledge with various perspectives, should be developed for better understanding learners’ fundamental literacy skills.

1. Introduction

Three main categories of vocabulary knowledge are form, meaning, and use (Nation, 2001). For basic vocabulary acquisition, understanding a word in its spoken form and/or written form, along with its meaning, is essential. It is noteworthy that in English-speaking countries (L1), writing with correct spelling is a reflection of good education. That is, spelling knowledge is not just the ability to write correctly spelled words but also reflects an expected level of mastery of the language for literate persons. One study reported the effects of spelling errors on the perception of writers (Kreiner, Schnakenberg, Green, Costello, & McClin, 2002), indicating that many spelling errors can affect how people (readers) perceive the intelligence of writers. From these points of view, considerable numbers of studies about English spelling acquisition have been conducted. By contrast, in countries where English is a foreign language (EFL), expectations regarding correct spelling of English words may be quite different.
In EFL countries, especially in Japan, spelling knowledge seems to be just one small aspect of learning English. However, in order to acquire English spellings, specific instruction is considered necessary (Allal, 2007). Learners who study English spelling without specific instruction often encounter serious problems in the acquisition process. However, adequate knowledge of the spelling acquisition process among Japanese EFL learners has not yet been developed. This is evident when looking at the very few studies focusing on Japanese EFL learners’ spelling acquisition. Moreover, spelling acquisition is one of the minor aspects in English language learning and teaching in Japan.

In studies of orthographic development that target beginner learners in English speaking countries, several sequential steps have been introduced (e.g., Ehri, 1995; Frith, 1985). A very simple summary of their claim starts from (a) recognizing the alphabet letters, (b) understanding the sound-letter correspondences, and then (c) gradually understanding orthographic rules. It is reasonable to assume that Japanese EFL learners’ spelling acquisition follows these sequential steps. In order to confirm these sequences, which is related to depth of vocabulary knowledge, conducting various types of tests might be appropriate.

Using various types of tests for same target items enables us to examine accurately and precisely what learners know and what they do not know. Nation and Webb (2011) summarized basic experimental design in vocabulary experiments, indicating that some experiments used more than one post-test to assess learners’ vocabulary retention. As summarized in their book, Webb’s (2005) study and Waring and Takaki’s (2003) study contained more than one test format for testing same words among Japanese EFL learners. Webb (2005) used ten tasks including receptive and productive formats for the same words (but nonsense words) to see whether the effects of different kinds of learning conditions result in gains of different parts of word knowledge (i.e., orthography, syntax, association, grammatical functions, and meaning and form). After this, Webb (2007) also used this methodology and conducted research using ten test formats to examine the effects of repetition on vocabulary knowledge. As he pointed out in his study, using both receptive and productive formats for the target words to compare receptive and productive vocabulary sizes. Using various types of carefully designed spelling test formats are necessary. Diagnostic spelling tests that focus on different orthographic skills would be helpful tools to achieve an “in depth understanding” among learners (e.g., Bruck & Waters, 1988; Kamhi & Hinton, 2000; Mori, 2007; Perin, 1982; Wade-Wooley & Siegel, 1997). In these studies, learners whose reading ability seemed sufficient were called good spellers (S+), and those without good spelling ability are called poor spellers (S-).

Some researchers have focused on discrepancies between reading ability and spelling ability instead of learning conditions. The ability to read aloud is still unclear. Some studies focused on the ability to translate words’ meanings and pronouncing words correctly are the different types of vocabulary knowledge. Moreover, need to translate words’ meanings and pronouns are different aspects of knowing a word. Therefore, to measure different types of vocabulary knowledge, specifically designed test formats are needed in order not to misunderstand the results from learners’ performances.

Similarly, Waring and Takaki (2003) tested same words (but nonsense words) with different test formats. They reported that various test formats resulted in different scores; for instance, a recognition test was easier than a multiple-choice test, and a multiple-choice test was much easier than a word translation test. Thus, even though the same target words were tested in different test formats, test-takers’ performance varied by the task differences. Some types of tasks are much more difficult than other types of tasks, which means multiple tests are needed to conduct a much more accurate and reliable assessment. However, the important thing to remember is to use equivalent test formats when comparing knowledge including multi-aspect knowledge. To solve this problem, Webb (2008) used equivalent receptive and productive test
formats for the target words to compare receptive and productive vocabulary sizes. Using equivalent test formats would provide precise information about learners’ acquisition situation.

In previous studies of diagnostic spelling tests, various types of spelling test formats were used. The test formats used in those studies can be classified into two general groups, such as recall tasks and recognition tasks. Most of the studies reported that recall tasks were much more difficult than recognition tasks. However, the differences between various task types have not been discussed properly, because most of the studies used both recall and recognition tasks in one research design with a variety of purposes, focused on several aspects related to depth of vocabulary knowledge.

Needless to say, translating words’ meanings, writing correctly spelled words, and pronouncing words correctly are the different types of vocabulary knowledge. Moreover, understanding words’ meaning, choosing correctly spelled words, and identifying correctly pronounced words are different aspects of knowing a word. Therefore, to measure different types of vocabulary knowledge, specifically designed test formats are needed in order not to misunderstand the results. Especially for the spelling test, some researchers have used a writing test (or essay writing) as one type of spelling test. However, learners are said to avoid some words with which they have no confidence while writing (Croft, 1982). Thus, a writing test cannot replace spelling tests. To know learners’ spelling knowledge about specific words, use of multiple types of carefully designed spelling test formats are necessary. Diagnostic spelling tests that focus on different orthographic skills would be helpful tools to achieve an “in-depth understanding” (Calhoon, Greenberg, & Hunter, 2010) of learners.

2. Background

2.1 Good and Poor Spellers

Some researchers have focused on discrepancies between reading ability and spelling ability among learners (e.g., Bruck & Waters, 1988; Kamhi & Hinton, 2000; Mori, 2007; Perin, 1982; Wade-Wooley & Siegel, 1997). In these studies, learners whose reading ability seemed sufficient are called good readers (R+), and those whose reading ability seemed insufficient are called poor readers (R-). Added to this, learners who have sufficient ability to correctly spell written English are called good spellers (S+), and those without good spelling ability are called poor spellers (S-). However, the definition of the ability to read is still unclear. Some studies focused on the ability to read aloud; however, some define the ability to read as reading comprehension. Thus, detailed consideration is needed to focus on these distinctions: read aloud or reading comprehension.

Given this situation, Takanami (2015) attempted to categorize learners according to eight types using three test formats focusing on the ability to decode, spell, and comprehend (Figure 1). Decoding is the ability to read aloud with comprehension, as shown by the correctly understanding the L1 meaning of the word. Thus, the eight patterns are as follows: (a) good
decoder, good speller, good comprehender (D+S+C+), (b) good decoder, poor speller, good comprehender (D+S-C+), (c) good decoder, good speller, poor comprehender (D+S+C-), (d) good decoder, poor speller, poor comprehender (D+S-C-), (e) poor decoder, good speller, good comprehender (D-S+C+), (f) poor decoder, poor speller, good comprehender (D-S-C+), (g) poor decoder, good speller, poor comprehender (D-S+C-), and (h) poor decoder, poor speller, poor comprehender (D-S-C-).

The results showed that if 70% accuracy is set as a sufficient level of acquisition, almost half of the learners were categorized as poor spellers (S-). Moreover, the results showed that there were almost no poor decoders (D-) and no poor comprehenders (C-) among Japanese university students. That is to say, the ability to write correctly spelled words is the key to diagnose learners’ deficiency with fundamental knowledge of individual words.

![Figure 1](image.png)

Figure 1. Eight categories of learners’ basic vocabulary knowledge: Decode, spell, and comprehend (Based on Takanami, 2015).

2.2 Summary of Previous Studies by Takanami (2014, 2015)

Almost all of the previous studies focusing on diagnostic spelling tests have targeted native speakers of English (L1). In other words, spelling knowledge among L2 learners or EFL learners had not been fully examined (Mori, 2007). Therefore, Takanami (2014, 2015) examined the effects of task differences such as different test formats to evaluate Japanese EFL learners’ spelling knowledge using high-frequency English words. In Takanami, 12 tests (Table 1), including six recall tasks and six recognition tasks, were developed in order to examine basic vocabulary knowledge based on Nation’s (2001, p. 347) suggestions of testing aspects of vocabulary knowledge. Among the test formats, six directions of vocabulary knowledge were examined focusing on spoken form (sound or pronunciation), written form (letter or spelling), and meaning (comprehension). The six directions are described in Figure 2, such as (a) L2 sound to L2 letter (S to L), (b) L2 letter to L2 sound (L to S), (c) L2 sound to L1 meaning (S to M), (d) L2 letter to L1 meaning (L to M), (e) L1 meaning to L2 sound (M to S), and (f) L1 meaning to L2 letter (M to L).
The results showed that three directions in recall tasks were difficult for Japanese EFL learners: that is, S to L, M to L, and M to S. Compared to these results, recognition tasks seemed too easy for the participants. In other words, Japanese university students were shown to have sufficient ability to recognize the basic aspects of vocabulary knowledge, such as sound (L2
spoken form: pronunciation), letter (L2 written form: spelling), and meaning (L1 meaning). The results also indicated that a specific test format in recognition tasks, the multiple-choice spelling test, was consistently reliable to assess learners’ spelling knowledge.

In Takanami (2014), a total of 35 high-frequency words were tested with the various types of test formats. After this, Takanami (2015) tested different 32 high-frequency words which were characterized by silent letters with the same test formats. In these studies, the method called implicational scaling (Hatch & Lazaraton, 1991) was utilized to confirm the difficulties of test formats. It is meaningful to compare the mean score of learners’ performance to examine task difficulties. Examining task difficulties is one of the most appropriate ways of understanding learners’ acquisition processes.

The results from Takanami (2014, 2015) succeeded in showing Japanese EFL learners’ weaknesses in spelling acquisition. However, some vowels (e.g., short vowel, long vowel, diphthong, and weak vowel), consonants, and sound-letter correspondences were not examined in the studies. In addition, some test formats need to be modified for the further research, because of those lower Cronbach’s alpha. Thus, this study focused on some other words to cover a broad range of characteristics in English words, and to upgrade and modify the test formats to confirm the findings from Takanami’s previous studies.

3. The Present Study

3.1 Purposes

The present study was not a simple replication study, but attempted to ascertain the findings from Takanami (2014, 2015). The purposes of this study are (a) to investigate learners’ weaknesses and strengths according to test performance, and (b) to reveal characteristics of good and poor spellers. A total of ten test formats were developed to examine the task difficulties. The test formats used in the study were based on the findings from the previous studies by the author (Takanami, 2014, 2015) in consideration of the internal consistency (Cronbach’s alpha). The present study intended to answer the following research questions (RQs).

RQ1: Do the results of multiple test formats illustrate learners’ weaknesses and strengths in fundamental literacy skills?

RQ2: What are the characteristics of good and poor spellers?

3.2 Method
3.2.1 Participants

Thirty-four Japanese university students participated in the study (ages 18 to 28). Some students majored in English while other students majored in economics or business economics, and their English proficiency levels varied. Instead of using proficiency test scores, such as TOEIC scores, this study utilized the written version of the Mochizuki Vocabulary Size Test.
(published in Aizawa and Mochizuki, 2010). This test was specifically designed to measure Japanese EFL learners’ estimated vocabulary size (lemmas), that is, the breadth of vocabulary knowledge. The word level of 1,000 to 4,000 were tested. The average score on the Mochizuki Test was about 2,300 \((M = 2,303.16, SD = 638.18)\). A total of 34 test-takers were divided into two groups: upper \((n = 13)\) and lower \((n = 21)\). The upper group’s vocabulary size was over the 2,500 word level \((M = 2,949.70, SD = 421.44)\), while the lower group’s vocabulary size was under the 2,500 word level \((M = 1,902.93, SD = 354.31)\). The statistical difference was significant between two groups and the effect size was large, \(t(32) = 7.788, p < .001, d = 2.75\).

### 3.2.2 Target Words

A total of 33 words were selected for the present study from junior high and high school textbooks’ corpus data (Chujo, Yoshimori, Hasegawa, Nishigaki, & Yamazaki, 2007). The target words in the present studies were not studied in Takanami (2014, 2015). By adding the present study’s target words, the series of studies enable to cover most of the characteristics in English words\(^2\), in terms of basic vowels and consonants. The words were summarized in Table 2, in reference to the list based on English high school textbooks.

<table>
<thead>
<tr>
<th>Target Word /PS/</th>
<th>Syl.</th>
<th>T_R</th>
<th>Target Word /PS/</th>
<th>Syl.</th>
<th>T_R</th>
<th>Target Word /PS/</th>
<th>Syl.</th>
<th>T_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>already /ˈɛrədi/</td>
<td>3</td>
<td>29</td>
<td>during /ˈdɔrn/</td>
<td>2</td>
<td>35</td>
<td>mouth /maʊθ/</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>another /əˈnʌðər/</td>
<td>3</td>
<td>33</td>
<td>enjoy /ˈen dʒɔɪ/</td>
<td>2</td>
<td>35</td>
<td>point /ˈpɔint/</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>ask /æsk/</td>
<td>1</td>
<td>35</td>
<td>fire /ˈfaɪ/</td>
<td>1</td>
<td>30</td>
<td>popular /ˈpɒpjəluə/</td>
<td>3</td>
<td>34</td>
</tr>
<tr>
<td>back /bæk/</td>
<td>1</td>
<td>35</td>
<td>flower /ˈflɔər/</td>
<td>2</td>
<td>26</td>
<td>put /pʊt/</td>
<td>1</td>
<td>34</td>
</tr>
<tr>
<td>boat /bɔt/</td>
<td>1</td>
<td>21</td>
<td>foot /fʊt/</td>
<td>1</td>
<td>27</td>
<td>sure /ʃʊər/</td>
<td>1</td>
<td>34</td>
</tr>
<tr>
<td>bring /brɪŋ/</td>
<td>1</td>
<td>34</td>
<td>forget /fərˈget/</td>
<td>2</td>
<td>31</td>
<td>thing /θɪŋ/</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>brother /ˈbrʌðər/</td>
<td>2</td>
<td>29</td>
<td>hour /ˈhaʊər/</td>
<td>1</td>
<td>33</td>
<td>thought /θɔːθ/</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>casual /ˈkæʒuəl/</td>
<td>3</td>
<td>–</td>
<td>however /həˈvaʊər/</td>
<td>3</td>
<td>34</td>
<td>tire /ˈtaɪər/</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>culture /ˈkələr/</td>
<td>2</td>
<td>31</td>
<td>love /laʊv/</td>
<td>1</td>
<td>35</td>
<td>travel /ˈtrævl/</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>dream /drɛm/</td>
<td>1</td>
<td>35</td>
<td>manual /ˈmænjuərəl/</td>
<td>3</td>
<td>–</td>
<td>true /tjuː/</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>drink /drɪŋk/</td>
<td>1</td>
<td>32</td>
<td>measure /meʒər/</td>
<td>2</td>
<td>17</td>
<td>with /wɪð/</td>
<td>1</td>
<td>35</td>
</tr>
</tbody>
</table>

*Note. "PS" = phonetic symbols. "T_R"= Textbook Range = Number of textbooks (which include each target word) out of 35 high school textbook series. "Syl."
= number of syllables. The symbol dash “—” means that the word was not in the list of Chujo, Yoshimori, Hasegawa, Nishigaki, & Yamazaki (2007).*

### 3.2.3 Test Formats (Dependent Measures)

In Takanami’s studies (2014, 2015), the results of six recall tasks were successful, showing that the tasks that required learners to write correctly spelled words were the most difficult, compared to a L2 to L1 translation task or a simple decoding task. However, in contrast to the results from recall tasks, some recognition tasks were too easy for the learners, showing a low value of Cronbach’s alpha. Thus, compared to the other tasks in the previous experiments, Test I (S to M) in the recognition task was deleted and integrated into Test J (M to S) in the recall task.
Test A (S to L) and Test C (L to S) were transformed into Test X (S⇔L) in recognition task. Therefore, among the six recognition formats, three test formats (Test E: S to M, Test G: L to M, Test K: M to L) remained and new version of integrated test (Test X: S⇔L) was added. Table 3 shows the test formats utilized in the present study.

Table 3

<table>
<thead>
<tr>
<th>Task type</th>
<th>Recognition</th>
<th>Recall</th>
<th>Aspects of vocabulary knowledge</th>
<th>Test format (Direction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>A</td>
<td>B (oral)</td>
<td>form</td>
<td>spoken L to L (L2 → L2)</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>L to S (L2 → L2)</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td></td>
<td>(spoken) form and meaning</td>
<td>S to M (L2 → L1)</td>
</tr>
<tr>
<td>G</td>
<td>H</td>
<td></td>
<td>(written) form and meaning</td>
<td>L to M (L2 → L1)</td>
</tr>
<tr>
<td>L</td>
<td>J (oral)</td>
<td></td>
<td>(spoken) form and meaning</td>
<td>M to S (L1 → L2)</td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td>(written) form and meaning</td>
<td>M to L (L1 → L2)</td>
</tr>
</tbody>
</table>

Note. “S” = sound (pronunciation), “L” = letter (spelling), and “M” = meaning (L1 translation). Test X is a combined version of Test A and Test C. Test formats in shaded section were deleted from the experiment.

It should be noted that Test K was traditionally used in the studies of diagnostic spelling tests. This test was in a multiple-choice four-option test format that did not include L1 translation. However, this test format can be categorized into “L1 meaning to L2 letter.” Figure 3 illustrates the ten test formats conducted in the present study.

**Figure 3.** Test formats in the present study (Modified version of Takanami, 2014, 2015)

### 3.2.4 Procedures

The study contained three sessions. First, in Session 1, the Mochizuki Test (published in Aizawa and Mochizuki, 2010) was conducted to assess participants’ estimated vocabulary size. The paper test version the Mochizuki Test is designed for Japanese EFL learners, and reliability and validity have been confirmed (e.g., Koizumi & Mochizuki, 2011). Second, in Session 2, six recall tasks were administered one week before the recognition tasks. Third, in Session 3, four
recognition tasks were conducted. In order to avoid giving clues to the test-takers from the materials, recall tasks should be tested before the recognition tasks. Figure 4 illustrates the procedures of three sessions in this study.

![Figure 4. Experimental procedures in the present study.](image)

3.2.4 Procedures
The study contained three sessions. First, in Session 1, the Mochizuki Test (published in Aizawa and Mochizuki, 2010) was conducted to assess participants’ estimated vocabulary size. The paper test version the Mochizuki Test is designed for Japanese EFL learners, and reliability and validity have been confirmed (e.g., Koizumi & Mochizuki, 2011). Second, in Session 2, six recall tasks were administered one week before the recognition tasks. Third, in Session 3, four recognition tasks were conducted. In order to avoid giving clues to the test-takers from the materials, recall tasks should be tested before the recognition tasks. Figure 4 illustrates the procedures of three sessions in this study.

3.2.5 Analysis of Task Performance
In order to analyze the task performance (test scores) in the present study, several statistical procedures were employed. Analyses were conducted independently for recall tasks and recognition tasks. First, two-way mixed ANOVA (analysis of variance) was conducted to compare the task difficulties to determine the sequential order. Second, in order to confirm the sequential development of task performance, implicational scaling (Hatch & Lazaraton, 1991) was used. Third, learners were categorized according to their performance on three tests (i.e., Test B: S to L, Test D: L to S, and Test H: L to M), which assessed the ability to (a) decode, (b) spell, and (c) comprehend, respectively. In this process, a sufficient level of acquisition was set as 70%.

3.2.6 Scoring
First, for the ten test formats, binary basis scoring (correct or incorrect) was used. However, in the L2 to L1 translation tasks, synonym words were considered as correct. Second, for the learners’ grouping, a cut-off point for 70% accuracy was adopted. Scores more than or equal to 70% accuracy of the total were considered as “pass”. Scores less than 70% accuracy of total were considered as “fail”. Results of pass or fail were transformed into binary numbers (1,0) for implicational scaling. Third, oral recall test (Test D) performances were evaluated by two researchers; discrepancies were reviewed and discussed until a consensus was reached.
4. Results and Discussion

4.1 Results of Task Difficulties

The analysis was conducted independently for recall tasks and recognition tasks. Descriptive statistics, reliability, and item discrimination for the ten test format (i.e., six recall tasks and four recognition tasks) are reported in Table 4.

Table 4
Descriptive Statistics, Reliability, Item Discrimination for the Ten Tests (N = 34)

<table>
<thead>
<tr>
<th>Test (k)</th>
<th>X (k = 33)</th>
<th>B (k = 33)</th>
<th>D (k = 33)</th>
<th>E (k = 33)</th>
<th>F (k = 33)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Task type)</td>
<td>(Recognition)</td>
<td>(Recall)</td>
<td>(Oral recall)</td>
<td>(Recognition)</td>
<td>(Recall)</td>
</tr>
<tr>
<td>Format</td>
<td>S ⇔ L</td>
<td>S to L</td>
<td>L to S</td>
<td>S to M</td>
<td>S to M</td>
</tr>
<tr>
<td>Group</td>
<td>Up</td>
<td>Low</td>
<td>Up</td>
<td>Low</td>
<td>Up</td>
</tr>
<tr>
<td>n</td>
<td>13</td>
<td>21</td>
<td>13</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>M</td>
<td>31.46</td>
<td>31.00</td>
<td>26.23</td>
<td>20.95</td>
<td>31.85</td>
</tr>
<tr>
<td>SD</td>
<td>2.76</td>
<td>1.67</td>
<td>4.62</td>
<td>5.40</td>
<td>1.21</td>
</tr>
<tr>
<td>Min</td>
<td>23</td>
<td>28</td>
<td>17</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>Max</td>
<td>33</td>
<td>33</td>
<td>31</td>
<td>28</td>
<td>33</td>
</tr>
<tr>
<td>α</td>
<td>.73</td>
<td>.87</td>
<td>.49</td>
<td>.86</td>
<td>.85</td>
</tr>
<tr>
<td>DIS</td>
<td>.22</td>
<td>.38</td>
<td>.07</td>
<td>.37</td>
<td>.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test (k)</th>
<th>G (k = 33)</th>
<th>H (k = 33)</th>
<th>J (k = 33)</th>
<th>K (k = 33)</th>
<th>L (k = 33)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Task type)</td>
<td>(Recognition)</td>
<td>(Recall)</td>
<td>(Oral recall)</td>
<td>(Letter recognition)</td>
<td>(Recall)</td>
</tr>
<tr>
<td>Format</td>
<td>L to M</td>
<td>L to M</td>
<td>M to S</td>
<td>M to L</td>
<td>M to L</td>
</tr>
<tr>
<td>Group</td>
<td>Up</td>
<td>Low</td>
<td>Up</td>
<td>Low</td>
<td>Up</td>
</tr>
<tr>
<td>n</td>
<td>13</td>
<td>21</td>
<td>13</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>M</td>
<td>31.31</td>
<td>27.48</td>
<td>31.15</td>
<td>26.95</td>
<td>22.15</td>
</tr>
<tr>
<td>SD</td>
<td>3.17</td>
<td>4.85</td>
<td>2.67</td>
<td>4.17</td>
<td>4.45</td>
</tr>
<tr>
<td>Min</td>
<td>22</td>
<td>13</td>
<td>24</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Max</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>α</td>
<td>.88</td>
<td>.86</td>
<td>.81</td>
<td>.73</td>
<td>.87</td>
</tr>
<tr>
<td>DIS</td>
<td>.41</td>
<td>.32</td>
<td>.30</td>
<td>.32</td>
<td>.42</td>
</tr>
</tbody>
</table>

Note. Up = Upper group, and Low = Lower group. Min = minimum value. Max = maximum value. α = Cronbach's alpha. DIS (i.e., item discrimination) was calculated using point-biserial correlations.

4.1.1 Characteristics of Recall Tasks

The results of two-way (2 [group: upper, lower] x 6 [test: B, D, F, H, J, L]) mixed ANOVA showed significant interaction between group and test, but the effect size was small, F(3.62, 115.87) = 4.125, p < .05, η^2 = .02. Subsequent analyses were conducted for multiple comparisons. First, simple main effect of task variance with the levels of each group was examined. Some significant task differences appeared with the upper group (all ps < .05), while some were not significant (all ps > .05). Using greater- or less-than symbols (<, >) or nearly equal symbol (≒) for the results of multiple comparisons, the task differences in the upper group can be illustrated as follows: J ≒ L < B ≤ F < H ≤ D (J < B, B < H, F < D; difficult → easy). Some significant task differences can be found with the lower group (all ps < .05), while some were not significant (all ps > .05). Using the symbols, the task differences in the lower group can be illustrated as follows: L ≒ J < B ≒ F < H < D (L < B, B < H and D; difficult → easy). Second, the simple
main effect of group variance with the levels of each task was examined. Using Bonferroni’s correction, significance was accepted at the $p < 0.05/8 = 0.00625$ level. The significant group difference appeared in Test H (L to M), and Test L (M to L; all $ps < .006$), showing that these two test formats succeeded in dividing upper group and lower group. Test B (S to L) and Test F (S to M) was marginally significant ($p = .006$). In contrast, Test D (L to S) and Test J (M to S) failed to divide participants with their estimated vocabulary size (all $ps > .006$). In other words, Test D (L to S) and Test J (M to S) had a low impact of spoiling learners’ judgement during the test.

In sum, Test H (L to M) and Test L (M to L) were difficult enough to examine learners’ basic vocabulary knowledge related to letter and meaning. Both for the upper group and lower group, the test formats that required learners to write correctly spelled words (Test B and Test L) seemed consistently difficult tasks through the three experiments (i.e., Takanami, 2014, 2015; the present study). The results of the internal consistency (Cronbach’s alpha) of each test showed low to high value (.49 to .88). The decoding task (Test D), which showed the lowest value, seemed too easy for the learners. In the present study, learners’ pronunciation on decoding task was graded with acceptance level by two researchers. Therefore, changing the scoring into strict criteria might have possibilities of producing different results on Test D.

### 4.1.2 Characteristics of Recognition Tasks

The results of two-way (2 [group: upper, lower] x 4 [test: X, E, G, K]) mixed ANOVA showed significant interaction between group and task variance for recognition tasks, and the effect size was large, $F(2.38, 76.28) = 5.374$, $p < .05$, $\eta^2 = .06$. The subsequent analyses were conducted for multiple comparisons. First, simple main effect of task variance with the upper group was not significant (all $ps > .05$), which means upper group learners could answer almost perfectly with recognition tasks. Only in the lower group were the differences between some tasks significant (all $ps < .05$). Expressed symbolically, the results were as follows: $E < G \approx K \approx X$ ($E < K$ and $X, G < X$; difficult $\rightarrow$ easy). Second, simple main effect of group variance with the levels of each task was examined. Using Bonferroni’s correction, significance was accepted at the $p < 0.05/6 = 0.0083$ level. The significant group difference only appeared in Test E (S to M; all $ps < .008$), showing that test performance of Test G (L to M), Test K (L2 letter recognition), and Test O (S $\leftrightarrow$ L) were not affected by the learners’ estimated vocabulary size (all $ps > .008$). However, the results of the internal consistency (Cronbach’s alpha) of each test showed high value (.73 to .88), which means that modification of recognition tasks from Takanami (2014, 2015) was successful.

In sum, for the learners whose vocabulary sizes were lower than the 2,500-word level, recognition tasks might be effective tools to confirm their achievement in basic vocabulary knowledge. The results were consistent compared to the findings from the previous research (Takanami, 2014, 2015), showing that recognition tasks were comparatively easy test formats for
Scalabilities were calculated for each target word. The sequential order of task difficulties was assumed from the raw scores of each recall test performance as follows (difficult → easy): J (M to S) < L (M to L) < B (S to L) < F (S to M) < H (L to M) < D (L to S). This ordering shows that Test J (M to S) was the most difficult (score poorly) and Test D (L to S) was the easiest (score highly) for the learners.

A total of ten words seemed scalable. Accuracy rate and Coefficient Scalabilities ($C_{scal}$) for each word are as follows: already (69% [.931]), ask (82% [.784]), back (84% [.688]), boat (76% [.689]), casual (58% [.756]), during (41% [.797]), measure (41% [.787]), point (91% [.895]), true (70% [.660]), and with (76% [.857]). The accuracy rates of ten words were varied ranging from low (41%) to high (84%). All these words showed lower scores on Test J (M to S), which is a productive (L1→L2) oral translation test. The sequential order for the recall tasks in three experiments (i.e., Takanami, 2014, 2015; the present study) was nearly identical. That is, Test B (S to L), Test J (M to S), and Test L (M to L) were consistently challenging for the learners; whereas, Test D (L to S), Test F (S to M), and Test H (L to M) were not so difficult. The difficult test formats found in the three experiments required learners’ productive knowledge of the English words. It might be effective to focus exclusively on these formats for further research.

4.2.2 Scalabilities in Recognition Tasks

Scalabilities were calculated for each target word. The sequential order of task difficulties was assumed from the raw scores of each recognition test performance as follows (difficult → easy): E (S to M) < G (L to M) < K (L: letter recognition) < X (S⇔L). This ordering shows that Test E seemed difficult compared to other tests (score poorly) and Test X was the easiest (score highly) for the learners. The results of this sequential order could not reflect difficulties of the test formats. However, results of no significant differences among recognition tasks were consistent with the previous results (Takanami, 2014, 2015).

Only three words (i.e., bring, sure, and thing) out of 33 words seemed to be valid for showing the sequential task difficulties, in terms of developmental sequence. The accuracy rates of three words were not so high among the target words; however Coefficient Scalabilities ($C_{scal}$) for each word were moderately high, such as bring (76% [.750] sure (77% [.677]), and thing (69% [.929]). All these scalable words showed lower scores on Test E (S to M) and Test G (L to M), related to the receptive (L2→L1) translation test. The scalable words in recognition tasks were different from the words in recall tasks. It seemed that most of the learners answered correctly among all recognition tasks, and the results indicated that there were considerable gaps
4.2 Results of Implicational Scaling

4.2.1 Scalabilities in Recall Tasks

Scalabilities were calculated for each target word. The sequential order of task difficulties was assumed from the raw scores of each recall test performance as follows: (difficult) J (M to S) < L (M to L) < B (S to L) < F (S to M) < H (L to M) < D (L to S). This ordering shows the sequential task difficulties, in terms of developmental sequence. The accuracy rates of three words were not so high among the target words; however, it can be said that learners who could answer correctly on the written translation spelling test (Test D: L to S), spelling test (Test B: S to L), and comprehension test (Test H: L2→L1 translation).

Added to the weak point of spelling ability (Test B: S to L), productive (L1→L2) translation tests (Test J: M to S, Test L: M to L) were difficult for the learners, which was consistent with the previous studies (Takanami, 2014, 2015). On the other hand, strong points were (a) the ability to connect sound and letter (Test X: S→L), (b) decoding test performance (Test D: L to S), (c) understanding the meanings of the written form of L2 words (Test H: L to M, Test G: L to M), and (d) recognizing correctly spelled words among distractors (Test K: four option multiple-choice).

4.3 Characteristics of Learners’ Test Performance

A total of 34 participants were categorized according to their test performance. It was striking that only seven learners (21%) could pass all test formats. The pass mark for the present study was set at 70% of accuracy for each item. Marking 70% of accuracy for already learned target words was not so difficult if the learners had sufficiently acquired the basic vocabulary knowledge. However, the results indicated that almost 80% of learners could not answer appropriately on ten test formats.

The category focusing on three fundamental skills showed some interesting results. On the three test formats (decode, spell, and comprehend), there were no poor decoders (D-), which was same tendency found in Takanami (2014, 2015). If the target words have been introduced in the past (junior-high or high schools), it was less likely to find D- (poor decoders). A total of 21 learners (62%) were categorized as D+S+C+ (good decoder, good speller, good comprehender) whose basic vocabulary knowledge (i.e., sound, letter, and meaning) reached a sufficient level. Ten learners (30%) were categorized as D+S-C+ (good decoder, poor speller, good comprehender), and three learners (8%) were categorized as D+S-C- (good decoder, poor speller, poor comprehender). Thus, a total of 13 learners (38%) were S- (poor spellers). Comparing the results with the multiple-choice spelling test (Test K: letter recognition), most of the learners could choose correctly spelled words among four options. Only one learner failed to answer the letter recognition test correctly, and this learner also failed to answer the written translation test (M to L). It can be said that learners who could answer correctly on the written translation spelling test (Test L: M to L) also could answer correctly on the decoding test (Test D: L to S), spelling test (Test B: S to L), and comprehension test (Test H: L2→L1 translation).

5. Conclusion

Alphabet letters are phonograms (a grapheme which represents a phoneme or phonemes). Spelling words in writing has a similarity to transcribing melodic patterns to musical notes. In other words, spelling errors might cause disharmony between writers and readers, in terms of
The results of the present study showed that multiple test formats illustrate learners’ weaknesses and strengths in fundamental literacy skills (RQ1). In the study, more than one-third (38%) of the Japanese EFL university students were still struggling with writing correct spellings of high-frequency English words. These learners were categorized as poor spellers (S-). Especially with the productive translation test formats, learners were more likely to produce incorrect answers. Even though the words were selected from basic high-frequency words from junior high and high school textbooks’ corpus data, the learners with a lower vocabulary size (fewer than 2,500 word levels) had difficulties in producing correctly spelled words.

Learners categorized as D+S-C+ (good decoder, poor speller, good comprehender) were also unsuccessful in the productive translation test (L1 meaning to L2 letter). However, strengths of the learners’ knowledge appeared from the test results. The learners in this study succeeded in showing high scores on (a) recognizing sound and letter connection, (b) decoding, (c) receptive (L2 → L1) translation, and (d) identifying correctly spelled words. The results also indicated that good spellers (S+) were always good decoders (D+) and good comprehenders (C+). On the other hand, poor spellers (S-) are basically not good at connecting sound and meaning, which is to say, poor spellers need to enhance the ability to automatically understand English sound and its meaning (RQ2). Added to this, poor comprehenders (C-) were always poor spellers (S-).

In sum, understanding words’ meaning and recognizing correctly spelled words are the first stage of vocabulary learning. It is reasonable to assume that orthographic knowledge of producing correct spelling of words will gradually improve through learning process, in the same manner as L1 learners. However, looking at the results from the study, more than one-third of university students still need to make an effort to achieve a sufficient level of spelling skill. Conducting multiple test formats will help remedial EFL learners who are still struggling with writing correct spellings of the English words. This process requires long-term research and effort both from learners and teachers (or researchers). However, the results from multiple test formats are likely to make a contribution to developing reliable test formats. Further research is needed targeting learners whose vocabulary sizes are higher than those in the present study. Also, using different target words could more effectively confirm this study’s findings. In addition, analyses of spelling errors should be conducted to focus on effective instructional methods for spelling acquisition among Japanese EFL learners.

Note

1. Implicational scaling (Hatch & Lazaraton, 1991) or the Guttmann procedure (scaling) is an effective analytical method to examine task difficulties including understanding the acquisition process. For the detailed explanation, see Hatch and Lazaraton (1991, pp. 204-222).

2. Characteristics of English words include short vowels (/i/, /e/, /æ/, /α/, /u/, and /u:/), long vowels (/iː/, /aː/, /æː/, /ʌ/, and /ʊ/), diphthongs (/eu/, /ou/, /au/, /oʊ/, /ɔɪ/, and /juː/), R-colored vowels
Characteristics of English words include short vowels (/iː/, /ɑː/, /æ/), weak vowels (/ʌ/, /ʊ/, /ɔ/, /ə/, /o/, and /əʊ/), and voiced or voiceless consonants (/θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /θ/, /dʒ/, /tʃ/, /dr/, /tr/, /dr/, /m/, /n/, /ŋ/, /l/, /r/, /j/, /w/)

3. Coefficient Scalabilities (Cscal) must be greater than .60 to be valid.

Acknowledgement

I would like to express my gratitude to the participants in the studies and Dr. Hideki Iimura for valuable comments and suggestions on an earlier version of this article and experiments.

References


Chujo, K., Yoshimori, T., Hasegawa, S., Nishigaki, C., & Yamazaki, A. (2007, June). The vocabulary of selected high school English textbooks (Issue Brief No. 40 Report B 71-92). Japan: College of Industrial Technology, Nihon University. Retrieved from http://www5d.biglobe.ne.jp/~chujo/data/Koukoutext.pdf#search='E9%AB%98%E7%AD%89%E5%AD%A6%E6%A0%A1%E8%8B%B1%E8%AA%9E%E6%95%99%E7%A7%91%E6%9B%B8%E3%81%AE%E8%AA%9E%E5%BD%99'


Lexical inferencing. In an experiment, a total of 35 Japanese EFL undergraduate and graduate students inferred the meanings of 11 target words whose prefixes were available for participants and 11 words whose prefixes were unavailable. All Japanese EFL learners needed to know a considerable amount of vocabulary size on EFL learners' lexical inferencing ability [Kreiner, Schnakenberg, Green, Costello, & McClint, 2002]. Effects of spelling errors on the perception of writers. The Journal of General Psychology, 129, 5-17. doi: 10.1080/00221300209602029


