Effects of Inference Question Sequence on Reading Comprehension

Maki SHIMIZU
Takasaki University of Health and Welfare

Abstract

The present study was implemented to find out, in order to help Japanese EFL learners comprehend passages better, in what order a series of inference questions should be arranged. In addition, the interaction between question sequence and learners’ L2 reading proficiency levels was examined. Inference questions were ranked from the viewpoint of the degree of proximity to a text theme, and questions asking for themes of entire passages were labeled as thematic questions. The main finding was that for lower L2 reading proficiency learners, the local-to-global transition, or the step-by-step way of questioning had a positive effect on their reading comprehension. In contrast, the learners of higher L2 reading proficiency showed a tendency to benefit from the opposite sequence. Furthermore, it was shown that these sequence effects were not limited to a particular type of question, but were spread over a series of questions.

1. Literature Review and Present Study

Teachers in EFL reading class often interpolate questions to check whether their students have comprehended materials properly, or to help their students understand materials properly. The first aim of the present study was to find out how we could more effectively pose inference questions in reading tasks; to be more specific, to learn in what order we should arrange a set of inference questions so that EFL learners’ comprehension would be most improved. I decided to confine my attention to the presentation sequence of questions. The second aim was to investigate the interaction between question sequence and L2 reading proficiency.

Fushimi’s (1992) research was conducted in the context of learning new concepts such as “metals,” and focused on establishing which sequence would be preferable in learning. Two Japanese passages were prepared, comprising about 600 and 2,100 characters. His research showed that the group given the explanation of the “well-known example” (e.g., “gold” in “metals”) first, and then the “misunderstood example,” defined as that learners understand incorrectly not to be an example of the denotation of the concept (e.g., “calcium” in “metals”), was able to learn better than the other group, which were given the explanation in the opposite order. What can be derived from this study is that when students learn something unfamiliar, a
presentation sequence that moves from easier to more difficult is more effective than the opposite sequence. Another researcher, Skinner (1968), developed the notion of small frames of instruction in the programmed instruction. In this process, a set of questions should be provided brick by brick by teaching machines or programmed sheets, and the change from one question to the next should be a short step so that anyone can understand each question. The same idea can be seen in Toyama and Ginbayashi (1975). They reconstructed computational systems in the order of “from general to specific.” What is common to these three previous studies is that the sequences are determined from easier items to more difficult items. Furthermore, considering their targeted students, this sequence seems to be more effective for poor learners or readers.

In contrast, in regard to good readers, the following assumptions can be made: (a) Good readers must have already understood the answers to easier questions by the time they are given a set of questions; therefore, an obvious effect of the sequence from easier to more difficult questions would not be observed, and (b) if good readers, on the other hand, are given a thematic question first, it may positively affect their comprehension of subsequent questions. This is because the thematic question will function as a guiding question which reveals an outline of a passage.

Thus, the following hypotheses were formulated for the present study.

Hypothesis 1. The question sequence in which learners answer inferential questions during the early stages of the sequence, far from thematic questions which are presented at the end of the sequence, is more beneficial for learners of lower L2 reading proficiency than the opposite sequence.

Hypothesis 2. The question sequence in which learners answer a thematic question during the early stages and answer inferential questions far from a thematic question during the final stages is more beneficial for learners of higher L2 reading proficiency than the opposite sequence.

The sequence in Hypothesis 1 is very general in reading comprehension tasks or reading comprehension tests. Also, note that the present study regards a question that asks for a main theme of a passage as a type of inference question, following Graesser, Singer, and Trabasso (1994). However in the following, I call it a thematic question, rather than an inference question.

2. Method

In all, 827 Japanese learners of EFL participated in this study. They were 10th-, 11th-, and 12th-grade students from four different public senior high schools in Japan. The 10th-grade students were all from prestigious schools in the local areas. They had completed at least English I (Ministry of Education, Science, Sports and Culture, 1999) in class. The 11th- and 12th-grade students were from another two schools, whose academic levels were average.

As with materials, two types of reading tests and one questionnaire were used in this study.
The reading test, adopted from Part A of the 6th section in the fiscal 2003 and 2004 English tests of the National Center for University Entrance Examinations (Daigaku Nyushi Center: DNC), was administered to examine the effects of question sequences on learners' reading comprehension (11 multiple choice questions with 4 choices). The narrative passages comprised 665 and 693 words, and had Flesch-Kincaid Grade Levels of 5.3 and 5.0, respectively. Another test was from parts of "3A" and "4B" in the STEP 2nd grade test 2 originally administered in 1997 (5 gap-filling questions and 5 question-and-answer questions, which had 4 choices). The STEP test was employed to measure the participants' L2 reading proficiency. A questionnaire was prepared to confirm that participants had not previously read the same passages.

Let us now look more closely at the two reading passages from the DNC tests. The five questions for the fiscal 2003 test and six questions\(^1\) for the fiscal 2004 test were organized starting with a question regarding detailed information, and moving to one seeking the main topic of the passage. Take an example from the fiscal 2003 test: (a) The first three questions--"According to Elena's parents, why would it be good for Elena to move to Japan?" (item 1), "What problems did Elena have after she entered junior high school?" (item 2), and "Why did Elena feel hurt when she was asked about going back to Brazil?" (item 3)--were found to be inference questions which would elicit from readers something like descriptions of characters' feelings or the reasons why a certain event occurred in a text; (b) the fourth question--"What did Natsumi mean when she said Elena is 'different'?"--was an inference question as well, but was more closely related to the theme of the passage than the previous three questions; and (c) the last question--"What is the main theme of this story?"--was a thematic question, which asked examinees to identify the main theme of the entire passage.

With regard to these DNC tests, two booklets were distributed to each participant. Booklet I included two reading passages, and Booklet II included 11 multiple-choice questions for the passages. The participants were instructed to make one choice from four options for each question, and mark the number in Booklet II directly. In other words, a sheet for marks only was not given. In Booklet II, one page was allocated for each question, and the participants were not allowed to go back to the earlier questions. The time to start and finish reading a passage, and to start and finish answering each question, was signaled by the administrators with a stopwatch.

During the research phases, the participants were randomly grouped into Group A or Group B. Then, according to the group, each participant received the two booklets, in one of which questions were ordered as shown in Table 1.

<table>
<thead>
<tr>
<th>Test</th>
<th>Group A (n = 403)</th>
<th>Group B (n = 391)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiscal 2003 Test</td>
<td>Items 1 → 2 → 3 → 4 → 5</td>
<td>Items 5 → 4 → 3 → 2 → 1</td>
</tr>
<tr>
<td>Fiscal 2004 Test</td>
<td>Items 1 → 2 → 3 → 4 → 5 → 6</td>
<td>Items 6 → 5 → 4 → 3 → 2 → 1</td>
</tr>
</tbody>
</table>

All the participants answered the fiscal 2003 test first, and after that, the fiscal 2004 test. As well
as a series of reading questions, information regarding key paragraphs was written on a question sheet. The purpose of these was to reduce the advantage of Group A in answering questions in the order of story plots.

The participants were given 13 minutes for the STEP tests, and 30.5 minutes for the DNC tests. For the DNC tests, the time for reading each passage (7 minutes each) and for answering each question (1.5 minutes each) was measured by test administrators, and the participants were obliged to follow their instructions. After completing the reading tests, a questionnaire was given.

To analyze the data, a two-way ANOVA was performed to assess the significance of the observed difference in means between the three different L2 reading proficiency groups (i.e., Lower, Middle, and Upper) and the two different groups in question sequence (i.e., Group A and Group B), and the interaction between them. The ANOVA was conducted with the scores of an interval scale, whereas Fisher’s exact tests were conducted item by item in the analysis of categorical data. As well as this, hierarchical cluster analyses were applied to investigate the effects of question order on each item. Dendrograms were examined making the following assumptions: (a) If there was no effect of question sequence on readers’ processes, the dendrograms of Group A and Group B would match perfectly because both groups had been given the same questions, and (b) if, on the contrary, there was an effect of question sequence on readers’ processes, the dendrograms of Group A and Group B would not match.

3. Results

3.1 STEP Test

In the STEP test, four items were left to make groups homogeneous with regard to L2 reading proficiency (Group A and Group B), and divide them into three different L2 reading proficiency groups (Lower, Middle, and Upper).

Table 2 Descriptive Statistics of STEP by Group and Result of T-Test

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>df</th>
<th>t</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Middle</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Upper</td>
<td>2.447</td>
<td>0.661</td>
<td>2.461</td>
<td>0.682</td>
<td>2.454</td>
<td>0.670</td>
<td>150</td>
<td>-0.121</td>
<td>-0.021</td>
</tr>
<tr>
<td>Total</td>
<td>0.856</td>
<td>0.935</td>
<td>0.844</td>
<td>0.960</td>
<td>3.154</td>
<td>1.774</td>
<td>792</td>
<td>0.180</td>
<td>0.130</td>
</tr>
</tbody>
</table>

Note. Full score = 4. g = effect size. .200 ≤ g (small) < .500; .500 ≤ g (medium) < .800; .800 ≤ g (large). Effect size of .199 and below is regarded as having no effect (Cohen, 1988).

A t-test confirmed that there was no significant difference between the mean scores of Group A and Group B, t (792) = 0.180, p > .05, and a one-way ANOVA and Tukey’s HSD post-hoc test revealed a difference among participants in the average scores, F (2, 794) = 3746.256, p < .001; that is, the “Upper” learners’ L2 reading proficiency was higher than that of the “Middle” learners.
and the “Middle” learners’ L2 reading proficiency was higher than that of the “Lower” learners. These results were also supported by effect sizes, $g$ and $\eta^2$.

### 3.2 Comparisons of Different Sequences of Answering Questions

The descriptive statistics of the DNC test\(^5\) can be summarized as follows (see also Table 3): (a) “Lower” in Group A obtained 5.536 ($SD = 1.876$), while “Lower” in Group B obtained 4.826 ($SD = 2.159$), (b) “Middle” in Group A obtained 5.824 ($SD = 2.263$), while “Middle” in Group B obtained 5.378 ($SD = 2.229$), and (c) “Upper” in Group A obtained 6.289 ($SD = 2.529$), while “Upper” in Group B obtained 6.803 ($SD = 2.332$).

<table>
<thead>
<tr>
<th>Group</th>
<th>Group A</th>
<th></th>
<th>Group B</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Mean</td>
<td>$SD$</td>
<td>n</td>
<td>Mean</td>
<td>$SD$</td>
</tr>
<tr>
<td>Lower</td>
<td>168</td>
<td>5.536</td>
<td>172</td>
<td>4.826</td>
<td>2.159</td>
</tr>
<tr>
<td>Middle</td>
<td>159</td>
<td>5.824</td>
<td>143</td>
<td>5.378</td>
<td>2.229</td>
</tr>
<tr>
<td>Upper</td>
<td>76</td>
<td>6.289</td>
<td>76</td>
<td>6.803</td>
<td>2.332</td>
</tr>
<tr>
<td>Total</td>
<td>403</td>
<td>5.792</td>
<td>391</td>
<td>5.412</td>
<td>2.330</td>
</tr>
</tbody>
</table>

A two-way (by question sequence and L2 reading proficiency) ANOVA was performed to assess the significance of the observed difference in means between the treatment and L2 reading proficiency, and the interaction between them. This revealed a significant main effect for the participants’ L2 reading proficiency, $F (2, 794) = 20.374$, $p < .001$, and a significant interaction between question sequence and L2 reading proficiency, $F (2, 794) = 4.152$, $p < .05$. In contrast, there was no effect for question sequence, $F (1, 794) = 1.676$, $ns$. The results of the effect size measured by $\eta^2$ were consistent with those of significance testing. Figure 1 shows mean scores by L2 proficiency group and sequence group. It should be noted that the two lines are crossed, which reveals an interaction.

![Figure 1. Mean scores of the DNC test by L2 reading proficiency and question sequence.](image-url)
To examine what patterns were found with regard to the interaction, one-way ANOVAs and t-tests using the Bonferroni approach, in which a significance level of .05/k (group numbers) is used for each individual test, were carried out. ANOVAs were performed by question sequence group to test the effect of the participants’ L2 reading proficiency level on scores, whereas t-tests were performed by L2 reading proficiency level to test the effect of question sequence on scores.

The results were as follows. With regard to Group A, on the one hand, the Brown Forsythe \( F^* \) test\(^6 \) showed no significant difference in mean scores between the three different L2 proficiency groups, \( F^* \) (2, 240.303) = 2.901, \( p = 0.057 \). Though it appeared that the higher L2 reading proficiency learners had, the higher the score they would receive, this was not statistically confirmed. Also, only a small effect size was obtained (\( \eta^2 = 0.016 \)). On the other hand, with regard to Group B, a significant effect for L2 reading proficiency level was revealed, \( F \) (2, 390) = 20.949, \( p < .001 \). By the post-hoc comparisons (the Tukey’s HSD tests), the differences between “Upper” and “Middle,” and “Upper” and “Lower” were shown at the .05 level. In addition, effect size supported these results, indicating a medium value of \( \eta^2 = 0.097 \).

Moreover, the t-tests were conducted repeatedly in order to examine the differences in the mean scores between Group A and Group B. A significant difference was only found between two groups of “Lower,” showing that the mean score of Group A was higher than that of Group B, \( t \) (338) = 3.235, \( p = 0.001 \). In contrast, both in “Upper” and “Middle” the difference between Group A and Group B did not reach significance (see Table 4).

Table 4 Results of T-Tests

<table>
<thead>
<tr>
<th>L2 Reading Proficiency Group</th>
<th>Group A</th>
<th></th>
<th>Group B</th>
<th></th>
<th>( t )</th>
<th>( df )</th>
<th>( **p &lt; .001 )</th>
<th>( g )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>168</td>
<td>5.536</td>
<td>1.876</td>
<td>172</td>
<td>4.826</td>
<td>2.159</td>
<td>338</td>
<td>3.235***</td>
</tr>
<tr>
<td>Middle</td>
<td>159</td>
<td>5.824</td>
<td>2.263</td>
<td>143</td>
<td>5.378</td>
<td>2.229</td>
<td>300</td>
<td>1.723</td>
</tr>
<tr>
<td>Upper</td>
<td>76</td>
<td>6.289</td>
<td>2.529</td>
<td>76</td>
<td>6.803</td>
<td>2.332</td>
<td>150</td>
<td>-1.300</td>
</tr>
</tbody>
</table>

To investigate what type of question was more likely to be affected by the question sequence, the number of participants who chose a correct or an incorrect answer were counted by group and item, and a Fisher’s exact test was run. The data of “Lower” and the data of “Upper” were targeted because the former revealed a significant difference and a small effect size, and the latter revealed a small effect size (see Table 4).

With regard to “Lower,” differences were found on four items; that is, three inference questions less related to the main theme of a passage (i.e., items 2 and 3 in the fiscal 2003 test, and item 3 in the fiscal 2004 test) and one thematic question (i.e., item 5 in the fiscal 2003 test). As to the “Upper” data, one coefficient reached a statistically significant level < .05. The question was one of the inferential questions not closely related to the main theme of the passage. Group A answered it first, whereas Group B answered it last. There were no differences between Group A and Group B in the other items (see Table 5).
Table 5 Results of Fisher's Exact Tests

<table>
<thead>
<tr>
<th>Group</th>
<th>Fiscal 2003 Test</th>
<th>Fiscal 2004 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Lower</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Correct</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>(n = 168)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Correct</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>(n = 172)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>56</td>
</tr>
</tbody>
</table>

Phi

Note. *p < .05. **p < .01. The last items (item 5 and item 6) were about the main theme of the whole text; the other items were inferential questions. In addition, the second last items (item 4 and item 5) were more strongly related to the main theme.

The results of cluster analyses are shown in the subsequent parts of this section. Figures 2 and 3 were made from the scores of “Lower,” while Figures 4 and 5 were made from the scores of “Upper.” Among these, only the dendrograms in Figure 4 proved coincident to each other. The other pairs did not agree.

Figure 2. The dendrograms of the “Lower” Group A’s (see the above dendrogram) and “Lower” Group B’s (see the below dendrogram) fiscal 2003 DNC test.

Figure 3. The dendrograms of the “Lower” Group A’s (see the above dendrogram) and “Lower” Group B’s (see the below dendrogram) fiscal 2004 DNC test.

Figure 4. The dendrograms of the “Upper” Group A’s (see the above dendrogram) and “Upper” Group B’s (see the below dendrogram) fiscal 2003 DNC test.

Figure 5. The dendrograms of the “Upper” Group A’s (see the above dendrogram) and “Upper” Group B’s (see the below dendrogram) fiscal 2004 DNC test.
4. Discussion

To summarize the results, there was no difference in means among L2 reading proficiency groups in Group A, but there were differences between “Upper” and “Middle,” and between “Upper” and “Lower” in Group B. These results were also supported by the comparisons between Group A and Group B in each of L2 reading proficiency groups, that is, “Lower” Group A obtained a higher mean score than “Lower” Group B, and “Upper” Group B showed a tendency to obtain a higher mean score than “Upper” Group A. Considering these results, the hypotheses in the present study can be answered as follows: Hypothesis 1 was supported, and Hypothesis 2 was partly supported.

These results can be interpreted from learners’ linguistic deficiency and from cognitive loads. The two perspectives will be discussed separately below. The first interpretation focuses on learners’ knowledge. It is considered that learners of lower L2 reading proficiency do not have enough vocabulary knowledge and grammatical knowledge to catch literal meanings, and that they are not able to make inferences even if they have the ability. The present local-to-global transition has functioned as a helpful guide for the learners to increase their inference making and understanding of literal meanings as well.

On the other hand, learners of higher L2 reading proficiency are considered to have adequate knowledge both of vocabulary and grammar, and are able to use their ability to make inferences. Therefore, the local-to-global sequence did not function as well as it did for learners of lower L2 reading proficiency. In contrast, the global-to-local sequence had small but positive effects on them. A thematic question placed before the other minor questions may have revealed a framework of the passages. This assumption is in line with previous studies which investigated the relationships between readers’ schemata and their understanding (e.g., Bransford & Johnson, 1973).

The second interpretation takes the same direction as that in Nassaji (2002). Nassaji discusses if a reader can construct mental representations at the textbase level effectively, their working memory capacity is not so much consumed with this lower-level processing, and is distributed among higher-level processing or inference generation processing. As to the present study, when the lower L2 reading proficiency learners looked at a question concerning a whole passage immediately after reading the passages, they may have already depleted their cognitive resources or working memory capacity in processing more basic units, and they did not have any more resources to process the information which was the subject of the question. Therefore, giving a cognitively demanding question to the lower L2 reading proficiency learners first had no effect, or rather, had a negative effect on their comprehension. In contrast, this was not the case with the higher L2 reading proficiency learners; they could afford to deliberate about the theme of an entire passage.

Finally, I would like to discuss the results of the Fisher’s exact tests and the cluster analyses.
in the remainder of this section. These statistical analyses were employed to examine on which item a positive or negative effect would particularly be found. The following findings resulted from the Fisher’s exact tests. First, there seemed no particular question type which had been directly affected by the question sequence. Results of “Lower” showed that the differences were found in three inferential questions slightly related to the central theme of the passages, and in one thematic question. The data of “Upper” showed that the difference between Group A and Group B was found in one inferential question scarcely related to the central theme of the passages. Second, considering the sequences of these five items, the effect of question order was likely to appear on the latter questions. It may take a certain time before the effect of question sequence starts to work positively. This does not contradict our common knowledge. With regard to cluster analyses, the two dendrograms in Figure 4 represented almost the same ones, but this was not the case with the pairs in Figures 2, 3, and 5. In sum, we could not find in the results of the cluster analysis any consistent pattern for the same question sequence groups.

5. Conclusion

One of the answers to how reading questions should be arranged, or how teachers should give a series of questions to their students in the classroom, was suggested by the present study. For poor readers, a particularly acceptable question sequence was that in which learners start at a question designed to tap some details written in a certain paragraph, and gradually come to questions designed to tap a main theme of an entire passage. This sequence is frequently used in class. In contrast, good readers showed the tendency to benefit more from the exact opposite sequence, that is, that during the early stages they are given a question asking for the main theme of an entire passage. From these findings, it can be concluded that there is an interaction between learners’ L2 reading proficiency and effective question sequences.

The major pedagogical implication of the present study is parallel to the viewpoint of “aptitude-treatment interaction” (Cronbach & Snow, 1977). That is, when instructors give a series of reading questions to their students in class, they should carefully devise their sequence while considering the students’ vocabulary, grammatical knowledge, and the cognitive load they are being subjected to. A small step sequence can be less demanding than the opposite sequence. In this, a set of questions are given, starting with questions which ask for details and moving to questions asking for the main theme of a passage. This sequence seems to be particularly effective for poor readers.

As for a suggestion for further research, if I do not limit my foci to inference questions, but shift my foci to all types of reading questions, I would be able to find a set of questions for this design more easily. For example, some questions ask for details explicitly written in texts, other questions ask for information elicited by integrating sentences, and furthermore, the remainder ask for the theme of a passage. These systematically arranged questions might be in some reading
textbooks for EFL or ESL learners. Furthermore, these questions are likely to appear more frequently in expository passages.

Notes

1. Because of the lack of a thematic question in the fiscal 2004 test, that question—“What is the main theme of this story?” (1. swimming, 2. friendship, 3. a rival, 4. competition)—was newly developed through a preliminary investigation, in which 40 high school students were involved.
2. These were confirmed to be appropriate for the present participants by conducting a pilot study, in which three other high school students voluntarily participated.
3. According to SPSS (1999), “clustering begins by finding the closest pair of objects (cases or variables) according to a distant measure, and combines them to form a cluster. The algorithm continues one step at a time, joining pairs of objects, pairs of clusters, or an object with a cluster, until all of the data are in one cluster” (p. 293). The present study used the Ward method, and the Squared Euclidean distance to measure intervals.
4. Based on Messick (1989, 1996), the test validity was examined from the structural aspect (considering Cronbach’s alpha) and the external aspect (considering the academic levels of the four schools). To summarize, mixed results were obtained, that is, positive evidence for the external aspect and relatively negative evidence for the structural aspect. In the end, six questions with low point biserial correlations (less than $r_{pb} = .200$) were dropped from the test.
5. The test validity of the DNC test was examined. A moderate coefficient, $a = .557$, was obtained for the structural aspect, and positive evidence was obtained for the external aspect.
6. This analysis was used because of the lack of homogeneity of variance.
7. It is defined as the relationship between a learner’s personal strengths and weaknesses in learning and the learning situation, including the type of program one is enrolled in.

References