Effects of reading subskills and test formats on reading performance

Masamichi MOCHIZUKI
Kazumi AIZAWA
Tetsuro FUJII
Atsushi IINO
Akiko KOCHIYAMA

Reading teachers instruct reading subskills such as skimming and scanning in order to help learners improve their reading ability. These subskills are also tested in reading comprehension tests. Research on reading, however, produces contradictory results about the number and relative difficulties of reading subskills. This study compared the difficulties of reading subskills in two test formats: multiple-choice and open-ended, and attempted to find the number of subskills that best explained reading ability by use of Structural Equation Modeling. The results found that the relative difficulties of reading subskills were inconsistent in multiple-choice and open-ended formats. The SEM analyses found that the two-skill model best explained reading comprehension ability in the open-ended format, which lends support to Song's (2008) finding, though no model fit in the multiple-choice format.

1. Introduction

Currently the model of reading as an interactive process prevails among L2 reading researchers (e.g. Eskey & Grabe, 1988). ESL textbooks such as Mikulecky (1990) include exercises designed to help learners develop reading subskills that facilitate top-down and bottom-up processing. The importance of developing these reading subskills can be seen in standardized English tests such as the *Test of English as a Foreign Language (TOEFL).* It continues to assess test-takers' reading comprehension by test items that are intended to measure reading subskills like comprension of a main idea, reference, and inference.

Questions arise from practicing teachers. Is it eventually effective to teach such reading subskills for total development of L2 reading? Are some subskills more important for L2 reading and worth more instruction and practice than others? Are some subskills more difficult to acquire and thus need more focus on them? This study addresses the issue of the contribution reading subskills make to reading ability.

2. Literature Review

Alderson (2000) reviews studies on reading skills and concludes that there is no agreement as to the construct of reading ability or the number of subskills, even among researchers who agree that reading is divisible. However, because we investigate the effects and ease of subskill instruction and acquisition, we adopt the view that reading ability is divisible into subskills.

Several studies address the relationships between item difficulties of reading comprehension questions and reading subskills but produce different results. Bensoussan, Sim, and Weiss in Alderson (2000) found that local questions were easier than global ones. Alderson also cites Kintsch and Yarbrough's findings that performance on macro-level process tasks, which dealt with global understanding, was always affected by poor rhetorical text organization, while performance on micro-level process tasks, which were related to local understanding, was not affected by rhetorical organization. This may suggest that local-level understanding is easier than global level understanding. These findings are partly supported by Ushiro, Nakagawa, Morimoto, Hijikata, Watanabe, and Kai (2008), who conducted two experiments to investigate the relationships between reading test formats and question types. In the first experiment, they employed open-ended questions that were converted from multiple-choice questions of A.C.E, GTEC, and TOEIC tests and found thematic questions the most difficult, inferential questions the second most, and paraphrase questions the least difficult. In the second experiment, they used the original multiple-choice questions of the above-mentioned tests and found no difference in item difficulty among the three question types. On the other hand, Aizawa, Yamazaki, Fujii, and Iino (2009) presented the opposite results as to relative difficulty of local- and global-level reading. They presumed eight subskills: main idea, skimming, scanning, local fact finding, global fact finding, local/global fact finding, cohesion and coherence, and inference. They examined the University Center Examination for item difficulty of these subskill question items and found that the most difficult question types were local fact finding, global fact finding, and inference in the order of difficulty, while the easiest types were main idea, cohesion and coherence. Thus, these studies present contradictory results as to the relative difficulty of subskill test items.

Song (2008) does not directly compare the relative difficulty of reading subskill question types but still she provides insights into this issue. Song investigated listening and reading comprehension test performances to see if the two skills might share common construct components since the two are similar as receptive skills. She compared three models of listening and reading comprehension: one-skill model (comprehension), two-skill model (explicit and implicit), and three-skill model (topic, detail, and inference). She found that the three-skill model best fit the listening test performance and the two-skill model fit the reading test performance. She ascribed the difference of the two receptive skill models to the degree of learner proficiency in the two skills: the learners were highly proficient in reading in English while they were less proficient in listening. They were able to score equally well on explicit questions in reading, global (topic), and local (detail), but differed in implicit question performance. Thus the two-skill model (explicit and implicit) best fit reading. On the other hand, the learners differed greatly in topic, detail, and inference question performances in listening, and thus the three-skill model (topic, detail, and inference) was adopted. This supports Alderson's (2000) argument that subskills more likely exist in beginning-level learners. Song's findings imply that reading skill can be subdivided into global, local, and inference subskills if learners' proficiency levels are rudimentary.

The literature we have reviewed presents rather inconsistent and complex views of reading subskills and the construct of L2 reading ability, which leads us to two intriguing issues. First, the difficulties of test items representing different reading subskills differed among studies (Bensoussan et al. cited in Alderson, 2000; Aizawa et al., 2009) and in different test formats even among the same study (Ushiro et al., 2008). Bensoussan et al. found local fact finding was easier than global fact finding, while Aizawa et al. found the opposite. Ushiro et al. found test items about the main idea were the most difficult, those on inference the second most and those on local information were the least difficult in the open-ended question format but did not find any difference in difficulty according to the different question types in the multiple-choice format. These findings intriguingly invite two research questions.

Research question 1: Are local fact finding questions easier than global fact finding ones?

Research question 2: Are item difficulties of different question types the same regardless of test formats?

The second issue is about the existence of reading subskills themselves. Song (2008) made an attempt to find how many skills comprise the construct of reading ability by means of Structural Equation Modeling. She found the two-skill model best fit the reading comprehension performance among the three models and the three-skill model best fit the listening comprehension performance. She attributed the adoption of different models for reading and listening to differences of the participants abilities of the two skills. The participants' reading ability was so advanced that there was no difference between their global and local question performances in reading, while their listening ability was less developed so that there was a difference between their global and local question performances in listening. That suggests, as Alderson (2000) argues, that learners' reading ability consists of more than two subskills when it is not highly developed. Will Song's finding apply to learners at different levels? This leads us to our third research question.

Research question 3: Does the two-skill model best fit reading comprehension?

This study addresses these research questions in order to improve instruction to help learners develop their reading ability.

3. Method

3.1 Materials and hypotheses

Two passages were chosen from a TOEFL Practice Test. One was a 329-word-long passage about American Indian tribes with Flesch-Kincaid Grade Level 8.1 and twelve multiple-choice questions. The other was a 334-word-long text about Marianne Moore, a poet, with Flesch-Kincaid Grade Level 8.8 and ten questions. Because the two passages were similar in length and readability, we regarded them as equal in terms of reading difficulty. We created open-ended comprehension questions by translating question stems into Japanese and removing the choices for the two passages; thus, there were two test formats for each passage. We then made a reading comprehension test that consisted of two passages with different formats (multiple-choice or open-ended).

The twenty-two questions were grouped into five question types: Topic, Detail, Reference, Vocabulary, and Inference, three of which were used in Song (2008). The types were defined as follows in this study. Topic questions challenge readers to understand explicit information related to the topic or main idea of a text. Detail questions challenge readers to understand information within a sentence. Reference questions ask readers to identify the referent of a personal pronoun. Vocabulary questions ask test-takers if they know the meaning of a word in context. Inference questions challenge test-takers to make inferences from what the text implies. Table 1 shows the categorization of the questions into five question types.

Table 1 *Numbers of Each Question Type*

Types	No. of Qs	Question Nos.
Topic	3	Indian 1 and 12. Moore 1.
Detail	8	Indian 2, 8, 9, 10, and 11. Moore 3, 5 and 7.
Reference	4	Indian 3 and 5. Moore 4 and 9.
Vocabulary	3	Indian 6. Moore 6 and 8.
Inference	4	Indian 4 and 7. Moore 2 and 10.

The multiple-choice (MC) questions were converted into open-ended (OE) ones by rewriting questions in Japanese and eliminating the choices. For example, the first question in the American Indian passage challenges readers to grasp the main idea of the passage, *i.e.*, a Topic question: “What does the passage mainly discuss?” This question was converted into “*Kore wa amerikan Indian no nanini tsuite kakareta bunsho desuka 10ji inaide* *kotaenasai*" (What does this passage tell about American Indians? Answer within ten letters.)

Care was taken so that test-takers would give similar responses to the ones in the MC questions. For instance, question item 7 in the American Indian passage is intended to measure readers' inference skills because the correct choice is not explicitly stated in the text and should be inferred: Question: "Which of the following is true of the Shoshone and Ute?"; Answer: "They were not as settled as the Hopi and Zuni." Thus, in converting this question into an OE one, the question should restrict the scope of inference to dwellings of the Shoshone and Ute so that test-takers would be able to come up with similar responses to the correct choice of the MC question. Thus, the OE question is "*Shoshoni zoku to yuto zoku wa donoyouna jukyoni sundeita to suisoku dekimasuka*" (What kind of dwellings do you infer the Shoshone and Ute lived in?)

In order to eliminate order and text effects, four versions of the test were created by combining the two test formats and passages and changing passage orders. Each version has 12 MC or OE questions about the American Indian passage and ten OE or MC questions about the Marianne Moore passage: Version A (Indian MC & More OE), Version B (Indian OE & More MC), Version C (Moore OE & Indian MC), and Version D (Moore MC & Indian OE). These questions types and test formats were analyzed to address Research questions 1 and 2.

In order to address Research question 3, the question types were grouped into larger categories. Song (2008) examined three models: one-skill (Comprehension), two-skill (Explicit and Implicit), and three-skill model (Topic, Detail, and Inference). We follow her categorization except for one-skill model. Song's two-skill model consists of Explicit and Implicit skills: here the Explicit skill consists of the Topic, Detail, Reference, and Vocabulary subskill question types, and the Implicit skill is the Inference question type. Song's three-skill model is made up of topic, detail, and inference skills: here, the topic skill refers to the topic question type, the detail skill is made up of detail, reference, and vocabulary question types, and the inference skill refers to the inference question type itself.

Thus, three hypotheses were put forwards:

Hypothesis 1: The proportion of correct answers to Detail question type items is higher than that of Topic question type items.

Hypothesis 2: The order of proportions correct of different reading subskill question type items is the same regardless of test formats (MC or OE).

Hypothesis 3: The two-skill model of reading ability best fits reading comprehension performance.

As to Hypothesis 3, we replaced Song's (2008) one-skill model with a five-skill model (Topic, Detail, Reference, Vocabulary, and Inference) and made an attempt to find the best model that would explain reading comprehension.

3.2 Procedure

From five public and private universities around the Kanto area, 202 students participated in the experiment. They differed in their English proficiency and major subjects such as medicine, education, economics, and foreign languages. In order to countervail the difference of English proficiency of the participants in the five universities and form a large one group of participants, the four versions of the reading comprehension test were randomly distributed to students in regular English classes in each university. They took the test in 40 minutes. The number of participants who took each version of the reading comprehension test was 51 (Versions A and B), 52 (Version C), and 48 (Version D).

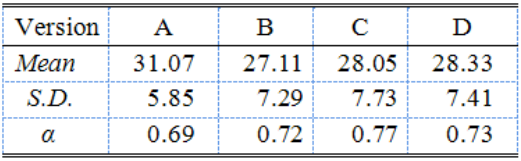
3.3 Scoring

Each question item was assigned two points. The multiple-choice questions were marked by one of us, and all the question items were given either 2 or 0 points. The open-ended questions were marked by two authors and all the questions were given either 2 (fully correct), 1 (partially correct), or 0 (incorrect). The final point for a participant for an open-ended question was decided on by averaging the two raters' scores for the item since the interrater reliabilities were high: A, .83; B, .93; C, .80; and D, .95.

4. Results

Seven participants scored zero points in either the first or the second passage and were excluded from further analyses, which reduced the numbers of participants who took the four versions to A (49), B (48), C (51) and D (47), 195 in total. The means of the four versions ranged between 27.11 (62%) and 31.07 (71%) and Cronback α between .69 and .77 as shown in Table 2.

Table 2 *Means and Cronback α of Four Versions*



Version	A	B	C	D
Mean	31.07	27.11	28.05	28.33
S.D.	5.85	7.29	7.73	7.41
α	0.69	0.72	0.77	0.73

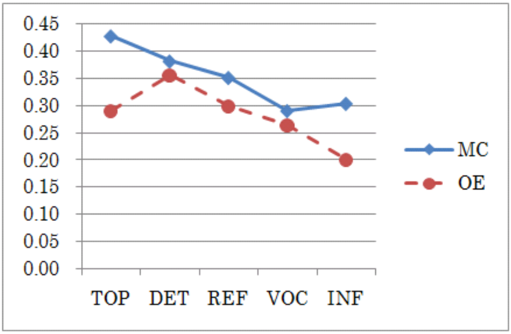


Figure 1 Proportions Correct of Question Types in MC and OE Formats

In order to test Hypotheses 1 and 2, we compared proportions correct of different question types which represented reading subskills. Since the numbers of different question type items differed from three to eight, proportions correct were calculated for each question type so that item difficulties could be compared among them (Figure 1).

The order of item difficulties is Vocabulary, Inference, Reference, Detail, and Topic for the MC format and Inference, Vocabulary, Topic, Reference, and Detail for the OE format. A two-way ANOVA found significant effects for question types and test formats and interaction ($F(4)=33.829$, $p<.001$; $F(1)=79.118$, $p<.001$; $F(4)=7.913$, $p<.001$). A Tukey test found that the mean of correct proportions to Topic was significantly higher than that of Detail in the MC format ($p<.05$), while the opposite was true in the OE format. These results partly support Hypothesis 1: the Detail question type items have a higher proportion correct than the Topic question type. Although the Detail question items were easier in the OE format, they were more difficult in the MC format than the topic questions. The results reject Hypothesis 2: The order of proportions correct of different reading subskill question type items is the same regardless of test formats (MC or OE).

In order to test Hypothesis 3, three models, two-skill (Explicit and Implicit), three-skill model (Topic, Detail, and Inference), and five-skill model (Topic, Detail, Reference, Vocabulary, and Inference) were analyzed for Structural Equation Modeling (SEM) to see the best fitting model. Each variable's normality was confirmed because the skewness and kurtosis values of each variable were within ±2.

Table 3 presents four fit indices: Comparative Fit Index (CFI), Goodness of Fit Index (GFI), Root Mean Square Error of Approximation (RMSEA), and Akaike's Information Criterion (AIC). CFI and GFI indices show that all the six models are well fit because they are all larger than 0.9 (Toyoda, 2007). However, only the two-skill and the five-skill models of the open-ended format may be considered to be fit because their RMSEA indices are smaller than 0.1, while those of the other models exceed 0.1. According to Toyoda, RMSEA indices should be below 0.05 if a model is considered to be fit and if values are between 0.06 and 0.1, a model may be regarded in a grey zone, which can be interpreted as either being fit or unfit.

Table 3 *Fit Indices of Models*

	Multiple-choice			Open-ended		
	2 skill	3 skill	5 skill	2 skill	3 skill	5 skill
CFI	0.91	0.924	0.91	0.934	0.931	0.934
GFI	0.957	0.967	-	0.975	0.975	-
RMSEA	0.128	0.131	0.128	0.093	0.107	0.093
AIC	40.873	39.381	50.873	33.417	34.885	43.417

Thus, only the two-skill and the five-skill models of the open-ended format have a possibility to be interpreted as being fit. Now we need to decide which model is better and if AIC indices play a part in it. The AIC indices of the two models indicate the two-skill model is better because its index is smaller than that of the five-skill model (Toyoda). This confirms Hypothesis 3.

5. Discussion

We now discuss three issues related to the three hypotheses. First of all, the results partly support Hypothesis 1: The proportion of correct responses to Detail question type items is higher than that of Topic questions type items. The results of the open-ended questions support the hypothesis, but those of the multiple-choice questions do not. The results showing Topic questions were easier than Detail questions in the multiple-choice format lend support to the findings of Aizawa et al. (2009).

However, the results of the open-ended format contradict those of Aizawa et al. and support the findings of Bensousan et al. cited in Alderson (2000). Furthermore, different orders of proportions correct in different testing formats support Ushiro et al. (2008). The fact that proportions correct of reading subskill questions in different testing formats may be attributed to low validity of test items that are intended to measure reading subskills. Although researchers regard some test items as measuring a certain reading subskill, they rarely check the validity of the test items as measuring the subskill. It is assumed that performance on the test items is affected not only by use of the subskill, but also other factors such as difficult words in the items and the relationship between the correct choice and distracters. For example, in the case of a multiple-choice format, a test item is more difficult when distracters are all related to the content of a text and are somehow similar to the correct choice than when they are not. Thus, when we collect data using a multiple-choice format test, we should make sure that the test items are valid in measuring the subskills we focus on. Furthermore, another factor we need to realize is even the same question type differs in difficulty as a reading comprehension question. For example, finding local facts varies in difficulty: some local facts are easier to find than others. Therefore, we should bear in mind that test items are valid in measuring certain reading subskills when we intend to compare their difficulty.

Second, we discuss reasons why the results do not support Hypothesis 2: The order of proportions correct of different reading subskill question type items is the same regardless of test formats (MC or OE). The results found that the difficulty order was Inference, Vocabulary, Topic and Reference, and Detail in the open-ended format, and that it was Inference and Vocabulary, Reference, Detail, and Topic in the multiple-choice format. They contradict Ushiro et al.'s (2008) findings. As we argued in the discussion of Hypothesis 1, it can be claimed that the difficulty of test items is affected not only by the use of a certain reading subskill, but also by other factors, so we should make every endeavor to make test items valid in measuring the characteristic we are interested in.

Third, the results partly support Hypothesis 3: The two-skill model of reading ability best fits reading comprehension performance. That is to say, the two-skill model best fits the reading comprehension performance in the open-ended format, but no model fits that of the multiple-choice format. Song (2008) ascribes the best fitting of the two-skill model in reading to no difference between the topic and the detail scores due to the high reading proficiency of the subjects. In this study, the mean of the reading comprehension performance was not high, between 62 and 71%, as shown in Table 2, but the means of Topic and Detail were close in the three-skill model (.29 and .36). On the other hand, the mean of Inference, .20, was substantially far below the means of the two subskills. That is why the two-model best fit the performance in the open-ended format. However, no explanation could be provided for the multiple-choice performances.

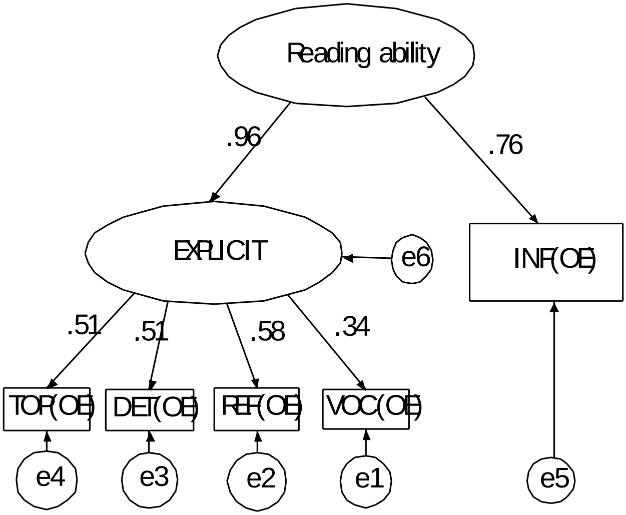


Figure 2 *Two-skill Model of the Open-Ended Format*
INF represents Implicit Knowledge.

Figure 2 illustrates the two-skill model of the open-ended format. It shows reading ability consists of explicit and implicit skills. The explicit skill is made up of subskills of Topic, Detail, Reference, and Vocabulary, whereas the implicit skill has only one subskill Inference. Even though the implicit skill has only one subskill, the standard estimation coefficient between reading ability and the skill is quite high (.76). The explicit skill has a very high coefficient (.96) and three subskills except for Vocabulary make similar contributions to it, with their coefficients ranging .51 to .58. The vocabulary subskill has a rather low coefficient (.34) and makes a small contribution to reading ability. This may be due to the fact that test-takers know the meaning of a target word but cannot understand the passage, or that they understand the text but do not know the meaning of the target word.

The study has three limitations. First, only two passages were used for the investigation. The results may have been affected by the genres of the texts. A wider range of genres needs to be employed in future studies. Second, the numbers of test items focused on subskills were not controlled as Table 1 shows: Topic 3; Detail 8; Reference 4; and Vocabulary 4. These numbers of test items should have been equalized. In future research vocabulary questions may be excluded because of their low contribution to reading ability so that other subskill questions can be increased in number. Third, as we have discussed above, factors that might have affected the difficulty of a test item were not controlled. In order to compare difficulties of test items focused on reading subskills, these factors should be controlled so that the test items may be valid in measuring the subskills in question. Because controlling the factors is extremely difficult, one idea would be to increase the number of test items in order to counterbalance the differences of difficulties of test items.

6. Conclusion

This study found that local fact finding questions were not always easier than global fact finding ones and item difficulties of different question types differed depending on test formats. It also found that the two-skill model best fit as the construct of reading ability in the open-ended format but no model fit in the multiple-choice format. These results might suggest that reading teachers do not have to worry about the relative importance of reading subskills when teaching them and focus on ones which they think are important. Future research considering the above-mentioned limitations may confirm the results the present study achieved or provide further insight into this issue.

Acknowledgements

This study is partially subsidized by a Grant-in-Aid for Scientific Research (C) (No. 19520514, Term 2007-2009). We really appreciate Dr. Rie Koizumi’s advice on the SEM analyses. Without her assistance this study could not have been accomplished though all errors are ours.

References

Aizawa, K., Yamazaki, A., Fujii, T., & Iino, A. (2009). The relationship between vocabulary knowledge and reading comprehension skills used on reading tests. *ARELE*, *20*, 111-120.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Eskey, D. E., & Grabe, W. (1988). Interactive models for second language reading: Perspectives on instruction. In P.L. Carrell, J. Devine, & D.E. Eskey (eds.) *Interactive approaches to second language reading*, pp.223-238. Cambridge: Cambridge University Press.

Mikulecky, B. S. (1990). *A short course in teaching reading skills*. Reading, MA: Addison-Wesley.

Song, M. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach.

Language Testing, *25*, 435-464.

Ushiro, Y., Nakagawa, C., Morimoto, Y., Hijikata, Y., Watanabe, F., & Kai, A. (2008). Effects of question types on item difficulty in two reading test formats: open-ended and multiple-choice. *ARELE, 19,* 201-210.

豊田秀樹編著 (2007). 『共分散構造分析[Amos 編]』東京：東京図書.