Method effects on reading comprehension test performance: text organization and response format

Miyoko Kobayashi *University of Warwick*

If tests are to provide an accurate measure of learners' language abilities, examiners must minimize the influence of intervening factors such as text organization and response format on test results. The purpose of this article is to investigate the effects of these two factors on second language learners' performance in reading comprehension tests. The study analyses the results of reading comprehension tests which were delivered to 754 Japanese university students. The main finding is that text organization and test format had a significant impact on the students' performance. When texts were clearly structured, the more proficient students achieved better results in summary writing and open-ended questions. By contrast the structure of the text made little difference to the performance of the less proficient students. This suggests that well-structured texts make it easier to differentiate between students with different levels of proficiency. Examiners have hitherto taken little notice of the impact of text structure and test format on students' results. By paying more attention to these factors, they will enhance the validity of their tests.

I Introduction

The theoretical framework for this article is twofold: research in the areas of language testing and of reading. In the area of language testing, Bachman's (1990) model of language ability (later revised in Bachman and Palmer, 1996) was the main inspiration for this study. By including 'method facets' as well as 'trait facets' in his discussion of language ability, Bachman draws our attention to a range of factors that can affect test performance and, therefore, jeopardize test validity. His model is the most influential and comprehensive available, although there has hitherto been little further research to validate it empirically.

According to Bachman, method facets can be divided into five categories:

1) testing environment;

Address for correspondence: Miyoko Kobayashi, CELTE, University of Warwick, Coventry CV4 7AL, UK; email: Miyoko.Kobayashi@warwick.ac.uk

Language Testing 2002 19 (2) 193–220 10.1191/0265532202lt227oa

2) test rubrics;
3) the nature of input;
4) the nature of the expected response; and
5) the interaction between input and response.

This study focuses on the third and fourth of these facets by manipulating text organization and test format, both of which play a significant role in reading comprehension tests.

1 Text organization

This study draws on insights from reading research to shed more light on the third of Bachman's categories: the 'nature of input'. A review of studies examining text characteristics and readability suggests that the coherence and organization of the text are significant factors influencing reading comprehension (Reder and Anderson, 1980; Davison and Kantor, 1982; Duffy and Kabance, 1982; Klare, 1985; Duffy *et al*., 1989; Olsen and Johnson, 1989; a series of studies by Beck and his colleagues, e.g., 1982, 1984, 1989, 1991, 1995). Surface-level features, such as syntactic or lexical elements, also affect readability but are secondary. Background knowledge is an important factor, but this has been extensively researched elsewhere (see, for example, Steffensen *et al*., 1979; Johnson, 1981; 1982; Carrell and Eisterhold, 1983; Alderson and Urquhart, 1983; 1985; 1988; Mohammed and Swales, 1984; Steffenson and Joag-Dev, 1984; Ulijn and Strother, 1990; Bernhardt, 1991; Salager-Meyer, 1991; Clapham, 1996).

There have been some attempts to characterize coherence by:

1) examining how sentences are related to one another (Connor, 1987; Olsen and Johnson, 1989; Connor and Farmer, 1990);
2) quantifying and mapping the links between key words and phrases (Hasan, 1984; Hoey, 1991); and
3) rating coherence holistically (Bamberg, 1984; Golden *et al*., 1988).

These attempts have given useful insights. However, some of them are too complicated for practical use, and none seem to characterize the concept of 'coherence' precisely enough for research purposes.

Various researchers have also tried to establish schemes to identify the overall structure of a text. Such schemes include:

- story grammar (e.g., Mandler, 1982);
- macro-structure (e.g., Kintsch and van Dijk, 1978);
- content structure analysis (Meyer, 1975a; 1985); and
- causal chains (e.g., Trabasso *et al.,* 1984).

Meyer's model of prose analysis seems to provide the most promising basis for research because it makes it possible to produce a content structure diagram showing the rhetorical relationships among the different parts of a text. The model helps show how these relationships account for the coherence of the text.

In Meyer's content structure analysis, idea units are organized in a hierarchical manner on the basis of their rhetorical relationships. The rhetorical relation at the highest level in the hierarchy is called the top-level rhetorical organization and this characterizes the text. The top-level rhetorical structure is identified as one of the following: 'collection', 'causation', 'response', 'description' and 'comparison' (Meyer later renamed 'response', calling it 'problem–solution', and this is the term I have used in this study). These five types of top-level relationships are thought to represent patterns in the way we think (Meyer, 1985: 20).

Meyer's first three text types ('collection', 'causation' and 'response') are on a continuum based on time and causality, as summarized in Figure 1. The link between the ideas is weakest in 'collection', where ideas are loosely associated with each other around a common topic. 'Time sequence' is another type of 'collection', for example when recounting events in chronological order. In the 'causation' relation, the ideas are related both in terms of time (i.e., one event happens before another) and causality (i.e., the earlier event causes the latter). Finally, the 'response' relation involves more interrelationship in ideas in that a solution is suggested in response to the existing causality.

The last two of Meyer's five categories ('comparison' and 'description') are on a different plane from the others because they are based on hierarchy or subordination of ideas. In a 'description' relation, ideas are arranged in a hierarchical manner: 'one argument is superordinate and the other modifies this superordinate argument' (Meyer, 1985: 20). The 'comparison' relation has at least two subordinate arguments which are linked by an element of comparison.

collection 1	A , B	only loosely-associated
collection 2	A then B	time sequence
causation	A ⇒ B	time sequence + causality (A: antecedent, B: consequent)
response	A ⇒ B ↑ C	time sequence, causality + response (solution in response to the cause)

Figure 1 Three types of rhetorical organization: collection, causation and response

This means that there is more interlinking in the 'comparison' relation than in the 'description' relation.

The five types of rhetorical organization represent the degree of interconnectedness of ideas, from loosely-organized to tightly-organized. If coherence is characterized as the degree of unity – i.e., how well a text holds together – then this classification helps identify the distinguishing features of coherent texts. Meyer and her associates suggest that a well-organized text would be better recalled, and a tight top-level rhetorical organization would enhance comprehension because the ideas in the text are closely interlinked (Meyer, 1975a; 1975b; Meyer *et al*., 1980; Meyer and Freedle, 1984; Meyer *et al*., 1993). Meyer and Freedle (1984: 125) suggest:

> This overlap in ideas covered may lead to more efficient storage in memory with more retrieval paths and resultant superior retention over time rather than retention of unrelated descriptions about a topic.

The Meyer model of text analysis has been applied by a great number of researchers (Kintsch and Yarbrough, 1982; McGee, 1982; Taylor and Samuels, 1983; Carrell, 1984; Richgels *et al*., 1987; Golden *et al*., 1988; Goh, 1990; Salager-Meyer, 1991). Their findings suggest that text organization has a significant effect on comprehension and that texts with a better or more natural structure enhance comprehension (also see Dixon *et al*., 1984; Urquhart, 1984). The present study builds on these findings and explores their applicability in foreign language reading comprehension tests.

My preliminary attempts to identify text types in naturally-occurring texts suggested that the 'comparison' text type could be regarded as an elaboration of the 'description' text type. Therefore, it was decided to modify Meyer's framework by combining 'description' and 'comparison' into a single category called 'description'. In addition, since too many text types would complicate the research design, it was decided to adopt only the first category of 'collection' as an example of the most loosely-organized text type: this was renamed 'association'. As noted above, Meyer herself renamed the 'response' text type, calling it 'problem–solution', which is a better indication of what it entails. Thus, this study investigated the comprehension of four types of top-level rhetorical organization: 'association', 'description, 'causation' and 'problem–solution'.

2 Response format

Returning to Bachman's model and concern with test format, it should be noted that Meyer and her associates used 'recall' as a way of measuring reading comprehension performance when examining the

effects of text structure on reading comprehension. This was true of other research studies on second language readers (e.g., Carrell, 1984; Urquhart, 1984). However, recall is normally associated with the compound of ideas in memory – a different issue from understanding – and it is not a common measure in the second/foreign language testing field. It therefore seemed more worthwhile and meaningful to examine text type effects with more conventional test formats. Text type effects in these formats would suggest that a test score might be partly an artefact of test format, and this would have important implications for language testers. Out of the available formats, cloze tests, open-ended questions and summary writing were selected, and the interaction between the response format and text type was examined.

The idea of exploring this area was prompted by several research studies which suggest that different test formats are measuring different aspects of language ability (see, for example, Reder and Anderson, 1980; Graesser *et al*., 1980; Kintsch and Yarbrough, 1982; Lewkowicz, 1983; Shohamy, 1984; Graves *et al*., 1991; Shohamy and Inbar, 1991). Among others, Kintsch and Yarbrough (1982) investigated the effects of two test formats – open-ended questions and cloze tests – on reading comprehension test performance. They suggest that open-ended questions can measure the reader's comprehension of main ideas of the text, whereas cloze tests will touch only upon local understanding and will not reflect the reader's overall comprehension. Since the impact of text organization on reading comprehension is the primary focus of the present study, their findings were particularly relevant to this study. It was therefore decided to adapt their approach in the present study.

Building on Kintsch and Yarbrough's research and the pilot study results (see below), this study added summary writing because this format was supposed to be even more sensitive to overall understanding than open-ended questions. According to Bensoussan and Kreindler (1990: 57), summary writing is 'a whole-text, super-macro-level skill'. Finally, it should be noted that this study did not examine the multiple-choice format, despite its popularity as a test format for assessing reading comprehension in a second/foreign language. The reason is that this format has a significant drawback in that test takers can guess the right answer without fully understanding the reading passage, and thus test validity is questionable (see, for example, Nevo, 1989; Katz *et al*., 1990; Royer, 1990; Weir, 1993).

II The purpose of the study

The objective of the study was to investigate whether two factors – text organization and response format – exercise a systematic influence on test results. If the measurement of reading comprehension is

unaffected by these factors, no obvious interaction is expected to emerge. On the other hand, if systematic interaction is observed, this suggests that text organization and/or test format will have a significant effect on reading-comprehension test performance.

It was also decided to include learners' language proficiency level as a third research variable following the findings of my preliminary study. Learners of different proficiency levels seemed to be affected by the other variables in different ways. The following null hypotheses were therefore formulated:

- Hypothesis 1: There is no interaction between reading-comprehension test performance and text organization and/or response format.
- Hypothesis 2: There is no interaction between reading-comprehension test performance, learners' language proficiency level, and text organization and/or response format.

III Methodology

1 Pilot study

The pilot study was conducted before the main study and involved 219 Japanese university students. Its purpose was, first, to examine the viability of the research questions and, secondly, to identify potential pitfalls in the proposed research methodology. To this end, the influence of a number of relevant variables was explored. These included: topic areas of reading passages, text length, text readability, the number of questions, the nature of questions, students' language proficiency and appropriacy of test level for the students. Although the variables were not tightly controlled, the findings suggested that text structure and response format had an important impact on reading comprehension. The main study was therefore designed to explore this further. In addition to this relatively large-scale pilot study, the preparation involved a series of mini-pilots and reviews by expert judges (see Section III.4 below) to ensure the quality of the test materials.

2 Participants

A total of 754 Japanese university students participated in the main study, the majority being 18–19 years of age and in the first or second years of their courses. All had previously had six years of English language learning at secondary schools. The students in intact English language classes were randomly divided into twelve groups, with each

student receiving one of a selection of reading comprehension tests (see Section III.5 below).

3 Materials

a English proficiency test In order to establish the comparability of the twelve groups, an English proficiency test, consisting of 50 multiple-choice grammar and vocabulary items, was conducted (for the relationship between knowledge of grammar and/or vocabulary and reading ability see, for example, Grabe, 1991; Alderson, 1993b). The test was designed to fit the level of the students in the light of the pilot study results. It drew on past papers of the Cambridge First Certificate and an English proficiency test for overseas students used in a British university. Statistical analysis confirmed that there was no significant difference between the twelve groups in their English language proficiency ($F = .39$ *d.f.* = 11, 723, n.s.). The test results were also used to divide the participants into three different proficiency groups according to the rank order of their scores – Low, Middle and High – as a basis for comparison at a later stage of the study. The test statistics were: $\bar{x} = 29.7$ out of 50; s.d.= 8.07; reliability alpha = .82; facility values ranging from .17 to .99 with a mean of .59; item-total correlation ranging from .08 to .53 with a mean of .34.

b Reading comprehension tests The texts used in the study were specially prepared to maximize control over the variables identified in the pilot study. Topic areas were first chosen and model texts were selected from several educational sources. Care was taken to minimize the potential effects of cultural bias or student familiarity with the topic (cf. Alderson and Urquhart, 1985; 1988; Clapham, 1996). Six topics were chosen and, for each topic, four different texts representing four text types were prepared, resulting in a total of 24 texts.

From the six topics, two sets of texts concerning 'international aid' and 'sea safety', were finally selected for use in the study on the basis of expert judgement (see Section III.4 below) regarding their suitability as representative samples of the selected text types. The mean length of the texts was 369.3 words (with the range of 352–384), and the mean score was 64.4 (with the range of 58.5–69.9) on the Flesch Reading Ease Formula, which is one of the most widely recognized readability indices.

After the eight texts had been selected, test items were developed for each text in three formats: cloze, open-ended questions and summary writing. The number of items for each test was:

- 25 for the cloze test;

- 5 for the open-ended question format (the greatest number of items that could plausibly be extracted from a text of this length); and
- 10 for summary writing (10 pieces of information were identified as key ideas to be included in the summary).

Two response formats – open-ended questions and summary writing – were set in Japanese, the students' first language, to eliminate undesirable effects of the use of English on reading performance.

The deletion rate (every 13th word) and starting points for deletion were decided on the basis of results of the pilot study and extensive analysis of potential cloze items (for details, see Kobayashi, 1995). It was also decided to avoid deleting proper nouns and numbers. When a deletion fell on these words, the subsequent word was deleted instead.

Expert judges were invited to analyse the items in detail in order to maximize the comparability of the test items across the eight texts (see Section III.4 below). In cloze tests and open-ended questions, for example, it was ensured that the numbers of different item types were consistent with one another across the tests. All the tests were then trialled with a small group of Japanese university students (n = 10), and some modifications were made in the range of item difficulty and wording of questions on the basis of this pilot result (see Appendix 1 for a sample test; for details, see also Kobayashi, 1995).

4 Expert judgement

Use of expert judgement is a fairly recent development in the second language testing field (e.g., Zuck and Zuck, 1984; Alderson and Lukmani, 1989; Alderson, 1993a; Cohen, 1993). In this study expert judges were asked to assist at different stages, ranging from text selection and item analysis to establishing marker reliability. Most of the judges had MAs in applied linguistics and were currently engaged in EFL teaching, materials development or testing consultancy. Where non-native speakers were involved, their English proficiency was of a sufficient level to enable them to study for higher degrees at British universities. Varying numbers of people were involved at different points. For example, four educated native speakers of English were asked to answer the cloze tests and to identify item characteristics; 10 educated native and non-native speakers of English were invited to identify and rate the importance of ideas in the texts to provide a basis for marking summaries, and so on.

Another example is text selection, which was conducted in the following manner: 27 people were given a description of the four text types adapted from Meyer (see Section I.1) and a set of 12 passages

(half of the 24 passages presented in a random order). They were asked to identify a text type for each passage. The majority identified the text types of more than two thirds of the passages as expected, whereas a small proportion of judges recognized the text types of less than half of the passages. This result showed that the text types were recognized by the majority of the judges in the same way as the researcher. It was decided to focus on the opinion of the judges who identified text types of at least two thirds of the passages. The passages that achieved a high level of agreement (more than 90%) were finally selected as representatives of the four text types.

5 Procedure

The test was administered by classroom teachers who were given detailed written directions. Written instructions were also prepared for students. Both sets of instructions were written in Japanese and piloted to minimize the risk of misunderstanding or confusion.

Ideally, all the participants would have received all the versions, thus facilitating comparison of test performance. However, this approach would have had two limitations. First, considering the length of time required, it would have been impractical for all the participants to take all 24 tests. This problem was overcome by increasing the sample size, thus securing groups that were comparable in language proficiency (see above). Secondly, the validity of the research would have been undermined if the participants had read the same or similar texts more than once. For the same reason, Shohamy (1984) questions the validity of a study by Samson (1983), who compared three test formats by allowing the participants to take several versions based on the same passage.

It was therefore decided that it would be most practical to give each student two tests: each one would have the same test format, and would be based on passages of the same text type, one from each of the two topics. This meant that there would be 12 participant groups, each taking a different set of test versions. For example, one group would take a cloze test with two Causation texts while another would write summaries of two Association texts. Table 1 summarizes the 12 participant groups. The 12 sets of tests were arranged so that each version would be randomly distributed among the participants. To eliminate an order effect, the order of the two texts in each set was counterbalanced. In the light of experience gained in the pilot study, the time for the test administration was set at 50 minutes (25 minutes for the proficiency test and 25 minutes for the reading test).

Table 1 Participant groups

Test format	Text type			
	Association	Causation	Description	Problem–solution
Cloze	Group 1 (n = 63)	Group 2 (n = 61)	Group 3 (n = 66)	Group 4 (n = 65)
Open-ended	Group 5 (n = 59)	Group 6 (n = 57)	Group 7 (n = 54)	Group 8 (n = 57)
Summary	Group 9 (n = 66)	Group 10 (n = 63)	Group 11 (n = 62)	Group 12 (n = 62)

6 Statistical analysis

The cloze tests were marked by the semantically and syntactically acceptable word scoring method. The results were analysed using SPSS/PC. For both the proficiency test and the reading comprehension tests, descriptive statistics (i.e., means, standard deviations, item-total correlations for individual items and reliability) were calculated. In addition, for the reading comprehension tests, the analyses included correlations with the proficiency test and *t*-tests. On the basis of the results of these initial statistics, ANOVAs (both one-way and two-way) were conducted to test the research hypotheses. The significance level was set at $p < .05$.

To assess the reliability of marking, 15% of the papers (n = 64) of open-ended questions and summary writing were independently marked by other expert judges (one other in open-ended questions and two others in summary writing) in addition to the researcher. All of these were native speakers of Japanese and experienced teachers of English, with MA degrees in TESOL from a British university. The correlations were .92 between the two markers for open-ended questions and between .85 and .90 among the three markers for summary writing.

IV Results

1 Overall results

On the whole, reliability values were higher in cloze tests regardless of text types (α = .86 ~ .90) in comparison to open-ended questions and summary writing (α = .69 ~ .79). However, this seemed to be because there were more items in the cloze test. When the values were adjusted by using the Spearman–Brown prophecy formula to

standardize the number of items across the different response formats (the number of items to be 50), the values for open-ended questions ($\alpha = .91 \sim .96$) and summary writing ($\alpha = .85 \sim .90$) rose dramatically. This showed that open-ended questions and summary writing could be as reliable as cloze tests.

The mean scores (converted in percentages) of the reading tests for four different types of text structure and three types of response format are shown in Figure 2 (see Appendix 2 for details of descriptive statistics).

In the cloze tests, the mean scores were highest in Association texts and lowest in Problem–solution texts. In other words, comprehension performance as measured by the cloze format was better in loosely-organized texts and became poorer as the text structure became tighter. This suggested that the presence of clear text structure did not help reading comprehension performance in cloze tests, and perhaps even hindered it. There are no other studies conducted in this area and it is difficult to explain this pattern. It may be related to the density of information: tightly-organized texts may compress more different ideas into a limited space so as to include all elements needed to develop an argument, and may therefore contain more new words (see Kintsch and Keenan, 1973). This is an interesting area to explore further.

On the other hand, in open-ended questions and summary writing, scores were lowest in Association texts, the most loosely organized texts. The highest scores were in Description texts in open-ended questions, and in Causation texts in summary writing. More generally, the two most tightly-organized texts (Causation and Problem–solution texts) produced the highest mean scores in summary writing, whereas equally high values were observed in three text types (Description, Causation and Problem–solution texts) in open-ended questions. This

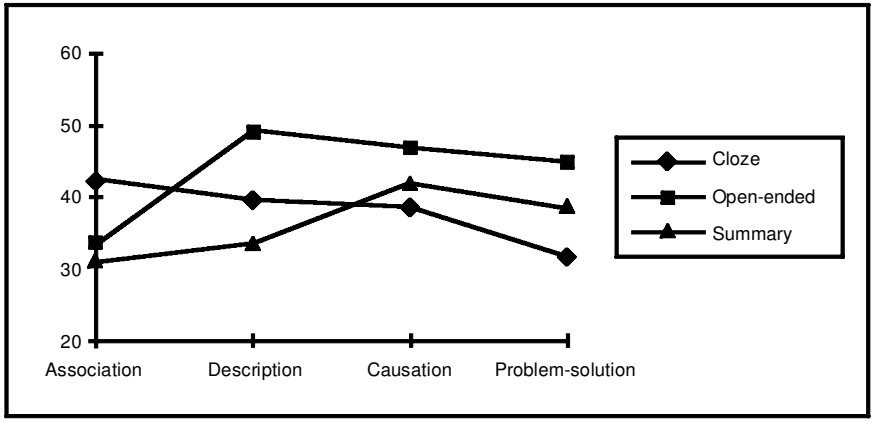

Figure 2 Comparison of text types: mean scores

may suggest that when reading comprehension is assessed through open-ended questions, it does not matter what kind of text structure is involved as long as there is some kind of structure.

To recapitulate, the results of this study suggest a clear distinction between cloze tests and the other two formats – i.e., open-ended questions and summary writing – in their interaction with types of text organization.

2 Hypothesis testing 1

To test the first research hypothesis, ANOVAs were conducted to examine the differences in the text type effects on reading-comprehension test performance. The results showed that the observed effects were statistically significant in all the three test formats individually and overall as shown in Table 2.

The effects of response formats were also examined for each text type, and it was confirmed that such effects were statistically significant in all text types except for the Causation texts (see Table 3). The response format effect in the four text types overall was also statistically significant.

The most important and interesting aspect of the results is that the two-way interaction between the two effects proved to be statistically significant (F (11, 723) = 6.149**, $p < .005$). This means that text

Table 2 Results of one-way ANOVA: effects of text organization

	F	*df*
Cloze	4.819**	3, 251
Open-ended	6.401**	2, 223
Summary	5.247**	3, 249
Overall	4.846**	3, 731

Note: **$p < .005$.

Table 3 Results of one-way ANOVA: effects of response format

	F	*df*
Association	8.644**	2, 185
Description	11.158**	2, 179
Causation	2.549	2, 178
Problem–solution	7.987**	2, 181
Overall	10.395**	2, 732

Note: **$p < .005$.

type and response format not only have significant effects on reading comprehension performance separately, but they also interact with each other. This confirms the statistical significance of the pattern shown in Figure 2.

The statistical results reported here clearly reject the first null hypothesis. That is, the differences in test performance observed across text types and response formats were statistically significant, and therefore it cannot be posited that test performance is unaffected by text type or response format.

3 Effects of learners' English proficiency level

When the results were examined by comparing three groups with varying English proficiency, further interesting findings emerged (see Figures 3–5; for details of descriptive statistics see also Appendix 3). In cloze tests (see Figure 3), higher proficiency learners performed consistently better than those with lower language proficiency, with more or less regular distances between them, even though the three proficiency groups varied in their test performance across different types of texts (especially the variation in the High and Low groups was significant) (see Table 5). This suggests that the distinction between different proficiency groups was clear regardless of the variation across text types within each group.

With open-ended questions (see Figure 4), the text-type effects varied according to the groups: the low group performed best with Description texts but there was little variation among the other three text types. By comparison, the two higher groups performed most poorly in Association texts but equally well in the other three text types. This suggests that in

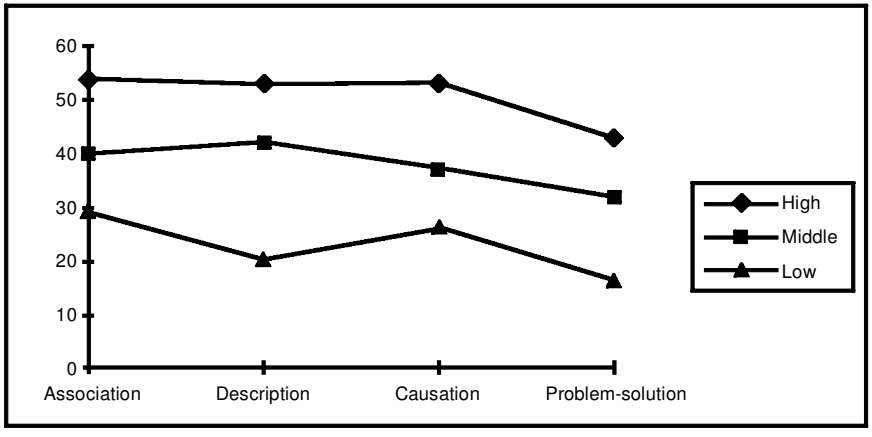

Figure 3 Cloze test results (as a percentage) by proficiency levels

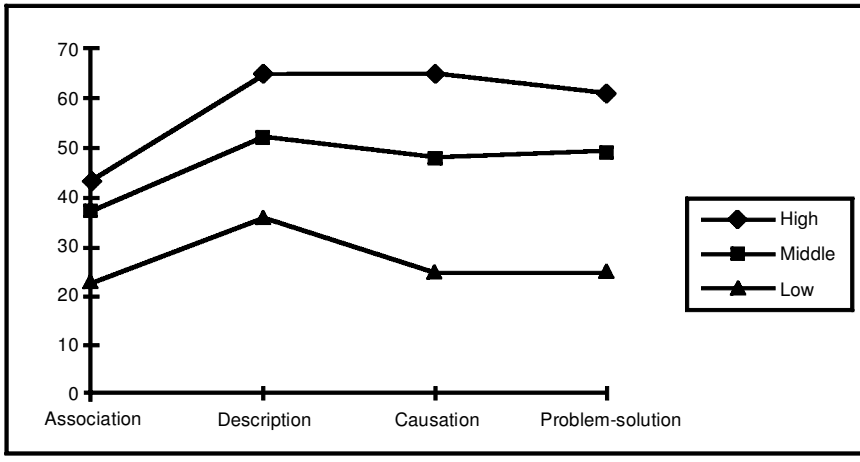

Figure 4 Open-ended questions results (as a percentage) by proficiency levels

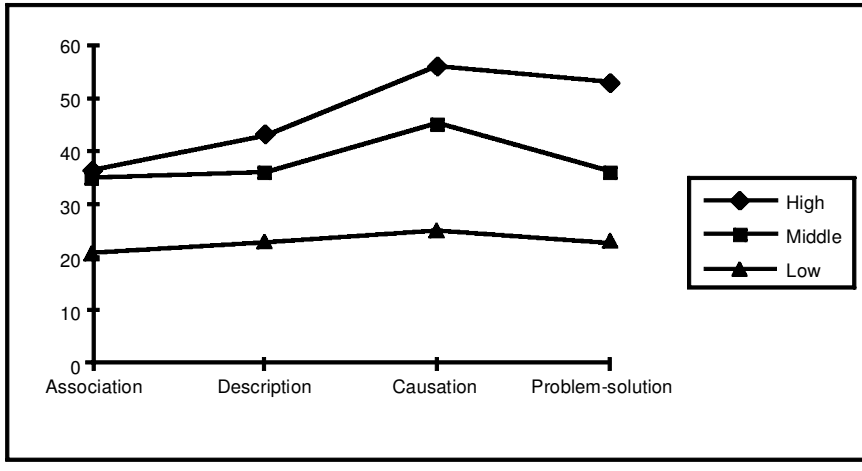

Figure 5 Summary writing results (as a percentage) by proficiency levels

open-ended questions higher proficiency learners will be disadvantaged if loosely-organised texts are used as an input for reading comprehension because they cannot demonstrate their ability fully.

In summary writing (see Figure 5), the diversity in the pattern across the three groups is striking. While the lowest group showed very little variation across text types, higher groups performed significantly better with more tightly-organized texts. This means that lower proficiency students could not exploit text structure when summarizing. By contrast, higher proficiency groups – who showed little difference from the lower proficiency students on the Association

texts – were better able to exploit text structure in line with their greater ability, and had their proficiency magnified. The impact of different kinds of text organization varied considerably across different proficiency groups, and this suggests that it is essential to take text types of reading passages into account when summary writing is to be used as a measure of reading comprehension.

4 Correlation with the proficiency test

The effect of learners' proficiency level was also examined by looking at the correlation coefficients between the results of the reading tests and the proficiency test (see Table 4). Although all the correlation coefficients were significant and the differences were small between different text types and response formats, the results shown here seem to confirm the findings presented in Section 3 above. For cloze tests, correlations with the proficiency test did not vary so much across text types, though Association texts tended to produce slightly lower correlations with the proficiency test.

By contrast, in open-ended questions and summary writing, results varied considerably depending on what type of text structure was involved. When texts with looser structures were used, the reading comprehension measured by these response formats did not correspond to general language proficiency as much as when more tightly-organized texts were used. This implies that students of higher proficiency may not be able to demonstrate their proficiency in reading comprehension tests if reading passages are not well structured. However, if structured texts are used, reading comprehension performance of students with higher proficiency will be proportionately better in open-ended questions and summary writing. This implies that reading comprehension performance will more accurately reflect learners' language proficiency when more structured texts are used with these test formats. In fact, in the Problem–solution texts, open-ended questions achieved the highest correlations with the proficiency test, even higher than cloze tests.

Table 4 Comparison of text types: correlation coefficients with the proficiency test

	Association	Description	Causation	Problem–solution
Cloze test	0.66**	0.78**	0.79**	0.75**
Open-ended	0.54**	0.65**	0.72**	0.81**
Summary	0.49**	0.56**	0.70**	0.75**

Note: $^{**}p < .001$.

5 Hypothesis testing 2

Table 5 summarizes the results of the one-way ANOVAs, which examined text type effects for the three proficiency groups. The table presents an interesting picture. First of all, F values were always highest for the High proficiency group, in all three response formats. This suggests that effects of text type were most evident for this group of learners, regardless of the response format. More interestingly, the difference between the proficiency groups became greatest in summary writing. Text type had no significant effect on the Low group, whereas its effect on the High group was greater than either of the other two response formats. This suggests that it does not matter for learners of lower language proficiency what kind of text structure is involved in the passages that are used as input for summary writing, but it does matter to a great extent for learners of higher proficiency. For open-ended questions, the difference between the proficiency groups was not so striking, but there was still a gradual increase in F values as the proficiency level rose. This suggests that the choice of passages was also important in this response format when learners' language proficiency is higher. By comparison, in cloze tests, both low and high groups showed significant values. However, as seen in Section 3 above, the significant text type effects were not so problematic in cloze tests because text types did not seem to affect discrimination between the different proficiency groups.

Table 6, which shows the results of one-way ANOVAs examining response format effects, presents an even more striking contrast between the High and Low proficiency groups. While the Low groups reached a significance level only in the Description text type, in the High group F values were significant in three text types and, moreover, the values were always greater. This suggests that response format effects are more evident when the learners' proficiency level is

Table 5 The results of analysis of variance: text type effects by proficiency levels

	Low (n = 239)	Middle (n = 238)	High (n = 258)
Cloze	4.25* ($d.f.$ = 3, 71)	n.s. ($d.f.$ = 3, 82)	5.34** ($d.f.$ = 3, 90)
Open-ended	2.80* ($d.f.$ = 3, 78)	3.24* ($d.f.$ = 3, 67)	6.94** ($d.f.$ = 3, 70)
Summary writing	n.s ($d.f.$ = 3, 78)	3.89* ($d.f.$ = 3, 77)	8.64** ($d.f.$ = 3, 86)
Total	n.s. ($d.f.$ = 3, 235)	3.82* ($d.f.$ = 3, 234)	6.61** ($d.f.$ = 3, 254)

Note: *$p < .05$, **$p < .005$.

Table 6 The results of analysis of variance: response format effects by proficiency levels

	Low (n = 239)	Middle (n = 238)	High (n = 258)
Association	n.s. ($d.f.$ = 2, 58)	n.s. ($d.f.$ = 2, 58)	8.95** ($d.f.$ = 2, 63)
Description	6.89** ($d.f.$ = 2, 58)	8.19** ($d.f.$ = 2, 59)	9.14** ($d.f.$ = 2, 56)
Causation	n.s. ($d.f.$ = 2, 56)	4.30* ($d.f.$ = 2, 60)	n.s. ($d.f.$ = 2, 56)
P-S	n.s. ($d.f.$ = 2, 55)	6.01* ($d.f.$ = 2, 49)	16.68** ($d.f.$ = 2, 71)
Total	n.s. ($d.f.$ = 2, 236)	8.57* ($d.f.$ = 2, 235)	14.28** ($d.f.$ = 2, 255)

Notes: *$p < .05$, **$p < .005$.

higher. In other words, learners with higher English language ability are more susceptible to different test formats and may not be able to demonstrate their reading comprehension skills when an unsuitable kind of format is used to measure them.

The results of two-way ANOVAs shown in Table 7 are again interesting. While both the main and interaction effects were obvious in the two higher groups, only the interaction effect attained a significance level in the Low group. A closer look at the F values also reveals an interesting tendency: the values were always higher in the High proficiency group. This suggests that the effects of these variables were greater for the learners of higher proficiency level of English. In other words, the combined choice of response formats and text structure of the passages was especially important for learners of higher proficiency.

The statistics presented here clearly indicate that the second null hypothesis with regard to learners' language proficiency should also

Table 7 The results of two-way analysis of variance by proficiency levels

	Low (n = 239)	Middle (n = 238)	High (n = 258)
Main effect			
Response format	n.s. ($d.f.$ = 2, 236)	8.57** ($d.f.$ = 2, 235)	14.28** ($d.f.$ = 2, 255)
Text type	n.s. ($d.f.$ = 3, 235)	3.82* ($d.f.$ = 3, 234)	8.87** ($d.f.$ = 3, 254)
Interaction effect	2.63* ($d.f.$ = 6, 227)	3.06* ($d.f.$ = 6, 226)	6.61** ($d.f.$ = 6, 246)

Notes: *$p < .05$, **$p < .005$.

be rejected. The examination has revealed complex interactions between this learner variable with text type and response format. The findings presented here confirm that text type and response format did affect reading comprehension performance of learners of different proficiency levels in different ways.

V Implications

1 Selection of reading passages

Typically, passages for reading comprehension tests have been arbitrarily selected without any coherent guiding principles. The basis of selection may be linguistic difficulty (e.g., vocabulary or syntactic complexity) or the tester's preferred topics. However, the findings of this study clearly suggest that it is essential to know in advance what type of text organization is involved in passages used for reading comprehension tests. Types of text organization do not seem to make much difference if reading comprehension is measured by cloze tests or if the learners' level of language proficiency is not high enough for them to be able to exploit text organization for comprehension. However, text types become most important if – as in summary writing – the test is intended to measure overall understanding. This is particularly significant with learners of higher language proficiency because they seem to be unfairly disadvantaged and their proficiency will not be reflected accurately in test performance when unstructured texts are presented. It is important that testing boards take these findings into account, especially since they may need to adjust test methods according to the test-takers' language proficiency levels.

It is of particular interest that learners of lower language proficiency did not benefit from clear text structure. At first sight this finding may seem surprising: clear structure ought to help them, and some research studies suggest that this is the case (e.g., Reder and Anderson, 1980). However, other studies suggest that better readers are more aware of overall text organization, and that this awareness enhances their comprehension (e.g., Meyer *et al*., 1980; Taylor and Samuels, 1983; Golden *et al*., 1988). The finding of the present study is in accordance with the second set of results.

This apparent contradiction in research findings may be related to the learners' level of language proficiency. That is, learners may need to have reached a certain proficiency level before being able to utilize text organization for overall understanding of the text. This seems to support a concept of linguistic threshold (e.g., Clarke, 1979; Alderson, 1984; Devine, 1988; Clapham, 1996; Ridgway, 1997). When the level of language proficiency is low, the learners have difficulty at a

decoding level, i.e., in word meanings and syntactic understanding. Consequently, they can rarely go beyond sentence-level meaning or literal understanding. It would be interesting to investigate the issue with learners of higher proficiency of English than those examined in this study.

2 Selection of test formats

Furthermore, different test formats seem to measure different aspects of reading comprehension. This has been already suggested in the literature (see Section I.2 above) and is confirmed by the present study. The results of this study further suggest that different test formats, or even different types of items within the same format, seem to measure different aspects of reading comprehension (for details, see Kobayashi, 1995). The finding is of particular significance to the test development process because it illuminates the complex nature of reading comprehension questions. It is highly recommended that language testers conduct in-depth qualitative analysis of test questions using expert judgement as in this study. This is particularly urgent in contexts where examination results have a strong impact on test-takers' lives – for example, university entrance exams.

Additional important findings are the complex interactions of test formats with text type and learners' language proficiency level. This result constitutes an important contribution to testing practice because the number of studies which have examined the relationship systematically is limited, especially in the second language field. The finding suggests that language testers ought not to choose formats of comprehension tests simply because the formats are familiar or convenient: they must be aware of the potential effects of test formats on reading comprehension performance. In this way, test scores will be more meaningful and reflect learners' reading ability more accurately. It is of the utmost importance for researchers to identify the exact nature of different test formats and make research findings accessible to testing practice.

VI Conclusions

1 Limitations of the study and future directions

This study has examined the test performance of a limited sample of Japanese university students, whose English language proficiency levels ranged from lower-intermediate to intermediate. It would therefore be interesting to replicate the study by extending learner variables, such as language proficiency level, age and different language

backgrounds. It would be particularly interesting to investigate whether the findings will apply to learners of higher language proficiency or even native speakers.

In this study, Meyer's (1975a) model of rhetorical organization was modified and four types of text organization were used as a research framework. As shown by the results of experts' judgement, the majority of judges identified types of text organization satisfactorily. This was important for the present research because text type was the main variable of interest. However, the four-type classification was by no means exhaustive or perfect. Meyer herself acknowledges the limitations of her classifications, and she has modified them over the years.

This raises further questions relating to the identification of text types. Can more agreement be achieved by training judges? Or is it impossible anyway? How can we handle naturally-occurring texts, which are often hybrid? There does not seem to be a clear answer to these questions. Sarig (1989: 86), in her text analysis, says that text types 'emerged', but in practice this is not so straightforward.

Nevertheless, the difficulty of text type identification should not prevent test developers from utilizing the findings of this study in their text selection. For practical testing purposes, it will be sufficient to decide whether a reading passage has some kind of structure or not, or whether some passages have clearer structure than others.

2 Final word

This research has employed Bachman's influential model of language ability as an organizing framework. The findings of this study have provided data supporting two aspects of his model: the nature of input and the nature of expected response. More research needs to be conducted in this area so that the findings reported here can be illuminated further, but the main implications are clear.

Test results are often used to make major decisions for educational purposes, and they play an essential role in many applied linguistics research projects. This study has clearly demonstrated that there is a systematic relationship between the students' test performance and the two variables examined: text type and response format. It is therefore vitally important for language testers, or anyone involved in assessment, to pay great attention to the test methods they use.

VII References

Alderson, J.C. 1984: Reading in a foreign language: a reading problem or a language problem? In Alderson, J.C. and Urquhart, A.H., editors, 1–27.

—— 1993a: Judgments in language testing. In Douglas, D. and Chapelle, C., editors, *A new decade of language testing research*. Alexandria, VA: TESOL, 46–57.

—— 1993b: Relationship between grammar and reading in an English for academic purposes test battery. In Douglas, D. and Chapelle, C., editors, *A new decade of language testing research*. Alexandria, VA: TESOL, 203–14.

Alderson, J.C. and **Lukmani, Y.** 1989: Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language* 5, 253–70.

Alderson, J.C. and **Urquhart, A.H.** 1983: The effect of student background discipline on comprehension: a pilot study. In Hughes, A. and Porter, D., editors, *Current developments in language testing*. London: Academic Press, 121–27.

—— editors, 1984: *Reading in a foreign language*. London: Longman.

—— 1985: The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2, 192–204.

—— 1988: This test is unfair: I'm not an economist. In Carrell *et al*., editors, 168–82.

Bachman, L.F. 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. and **Palmer, A.S.** 1996: *Language testing in practice*. Oxford: Oxford University Press.

Bamberg, B. 1984: Assessing coherence: a reanalysis of essays written for the national assessment of educational progress, 1969–1979. *Research in the Teaching of English* 18, 305–19.

Beck, I.L., McKeown, M.G. and **Gromoll, E.W.** 1989: Learning from social studies texts. *Cognition and Instruction* 6, 99–158.

Beck, I.L., McKeown, M.G., Omanson, R.C. and **Pople, M.** 1984: Improving the comprehensibility of stories: the effects of revisions that improve coherence. *Reading Research Quarterly* 19, 263–77.

Beck, I.L., McKeown, M.G., Sinatra, G.M. and **Loxterman, J.A.** 1991: Revising social studies text from a text-processing perspective: evidence of improved comprehensibility. *Reading Research Quarterly* 26, 251–76.

Beck, I.L., McKeown, M.G. and **Worthy, J**. 1995: Giving a text voice can improve students' understanding. *Reading Research Quarterly* 30, 220–38.

Beck, I.L., Omanson, R.C. and **McKeown, M.G.** 1982: An instructional redesign of reading lessons: effects on comprehension. *Reading Research Quarterly* 17, 462–81.

Bensoussan, M. and **Kreindler, I.** 1990: Improving advanced reading comprehension in a foreign language: summaries vs. short-answer questions. *Journal of Research in Reading* 13, 55–68.

Bernhardt, E.B. 1991: *Reading development in a second language: theoretical, empirical and classroom perspectives*. Norwood, NJ: Ablex.

Carrell, P.L. 1984: The effects of rhetorical organization on ESL readers. *TESOL Quarterly* 18, 441–69.

Carrell, P.L., Devine, J. and **Eskey, D.,** editors, 1988: *Interactive approaches to second language reading.* Cambridge: Cambridge University Press.

Carrell, P.L. and **Eisterhold, J.C.** 1983: Schema theory and ESL reading pedagogy. *TESOL Quarterly* 17, 553–73.

Clapham, C. 1996: *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Clarke, M. 1979: Reading in Spanish and English: evidence from adult ESL students. *Language Learning* 29, 121–50.

Cohen, A.D. 1993: The role of instructions in testing summarizing ability. In Douglas, D. and Chapelle, C., editors, *A new decade of language testing research*. Alexandria, VA: TESOL, 132–60.

Connor, U. 1987: Research frontiers in writing analysis. *TESOL Quarterly* 21, 677–96.

Connor, U. and **Farmer, M.** 1990: The teaching of topical structure analysis as a revision strategy for ESL writers. In Kroll, B., editor, *Second language writing*. Cambridge: Cambridge University Press, 126–39.

Davison, A. and **Kantor, R.** 1982: On the failure of readability formulas to define readable texts: a case study from adaptations. *Reading Research Quarterly* 17, 187–209.

Devine, J. 1988: The relationship between general language competence and second language proficiency: implications for teaching. In Carrell *et al*., editors, 262–77.

Dixon, R.A., Hultsch, D.F., Sinon, E.W. and **van Eye, A.** 1984: Verbal ability and text structure effects on adult age differences in text recall. *Journal of Verbal Learning and Verbal Behavior* 23, 569–78.

Duffy, T.M., Higgins, L., Mehlenbacher, B., Cochran, C., Wallace, D., Hill, C. Haugen, D., McCaffrey, M., Burnett, R., Sloan, S. and **Smith, S.** 1989: Models for the design of instructional text. *Reading Research Quarterly* 24, 434–57.

Duffy, T.M. and **Kabance, P.** 1982: Testing a readable writing approach to text revision. *Journal of Educational Psychology* 74, 733–48.

Goh, Soo Tian 1990: The effects of rhetorical organization in expository prose on ESL readers in Singapore. *RELC Journal* 21, 1–13.

Golden, J., Haslett, B. and **Gauntt, H.** 1988: Structure and content in eighth-graders' summary essays. *Discourse Processes* 11, 139–62.

Grabe, W. 1991: Current developments in second language reading research. *TESOL Quarterly* 25, 375–406.

Graesser, A.C., Hoffman, N.L. and **Clark, L.F.** 1980: Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior* 19, 135–51.

Graves, M.F., Prenn, M.C., Earle, J., Thompson, M., Johnson, V. and **Slater, W.H.** 1991: Improving instructional texts: some lessons learned. *Reading Research Quarterly* 26, 110–32.

Hasan, R. 1984: Coherence and cohesive harmony. In Flood, J., editor, *Reading Comprehension.* Newark, DE: International Reading Association, 191–219.

Hoey, M. 1991: *Patterns of lexis in text*. Oxford: Oxford University Press.
Johnson, P. 1981: Effects on reading comprehension of language complexity and cultural background of a text. *TESOL Quarterly* 15, 169–81.
—— 1982: Effects of reading comprehension of building background knowledge. *TESOL Quarterly* 16, 503–16.
Katz, S., Lautenschlager, G., Blackburn, A. and **Harris, F.** 1990: Answering reading comprehension items without passages on the SAT. *Psychological Science* 1, 122–27.
Kintsch, W. and **Keenan, J.** 1973: Reading rate as a function of number of propositions in the base structure of sentences. *Cognitive Psychology* 6, 257–74.
Kintsch, W. and **van Dijk, T.A.** 1978: Toward a model of text comprehension and production. *Psychological Review* 85, 363–94.
Kintsch, W. and **Yarbrough, J.C.** 1982: Role of rhetorical structure in text comprehension. *Journal of Educational Psychology* 74, 828–34.
Klare, G.R. 1985: *How to write readable English*. London: Hutchinson.
Kobayashi, M. 1995: Effects of text organisation and test format on reading comprehension test performance. Unpublished PhD Thesis, Thames Valley University, London.
Lewkowicz, J.A. 1983: Method effect on testing reading comprehension: a comparison of three methods. Unpublished MA Thesis, University of Lancaster.
Mandler, J.M. 1982: Some uses and abuses of a story grammar. *Discourse Processes* 5, 305–18.
McGee, Lea M. 1982: Awareness of text structure: effects on children's recall of expository text. *Reading Research Quarterly* 17, 581–90.
Meyer, B.J.F. 1975a: *The organization of prose and its effects on memory*. Amsterdam: North-Holland.
—— 1975b: Identification of the structure of prose and its implications for the study of reading and memory. *Journal of Reading Behaviour* 7, 7–47.
—— 1985. Prose analysis: purpose, procedures, and problems: parts I and II. In Britton, B. and Black, J.B., editors, *Understanding expository text*. Hillsdale, NJ: Lawrence Erlbaum, 11–64, 269–304.
Meyer, B.J.F., Brandt, D.M. and **Bluth, G.J.** 1980: Use of top-level structure in text: key for reading comprehension of ninth-grade students. *Reading Research Quarterly* 16, 72–103.
Meyer, B.J.F. and **Freedle, R.O.** 1984: Effects of discourse type on recall. *American Educational Research Journal* 21, 121–43.
Meyer, B.J.F., Marsiske, M. and **Willis, S.L.** 1993: Text processing variables predict the readability of everyday documents read by older adults. *Reading Research Quarterly* 28, 234–49.
Mohammed, M.A.H. and **Swales, J.M.** 1984: Factors affecting the successful reading of technical instructions. *Reading in a Foreign Language* 2, 206–17.
Nevo, N. 1989: Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing* 6, 199–215.

Olsen, L.A. and **Johnson, R.** 1989: A discourse-based approach to the assessment of readability. *Linguistics and Education* 1, 207–31.

Reder, L.M. and **Anderson, J.R.** 1980: A comparison of texts and their summaries: memorial consequences. *Journal of Verbal Learning and Verbal Behavior* 19, 121–34.

Richgels, D.J., McGee, L.M., Loman, R.G. and **Sheard, C.** 1987: Awareness of four text structures: effects on recall of expository text. *Reading Research Quarterly* 22, 177–96.

Ridgway, T. 1997: Thresholds of the background knowledge effect in foreign language reading. *Reading in a Foreign Language* 11, 151–68.

Royer, J. 1990: The sentence verification technique: a new direction in the assessment of reading comprehension. In Legg, S. and Algina, J., editors, *Cognitive assessment of language and math outcomes.* Norwood, NJ: Ablex.

Salager-Meyer, F. 1991: Reading expository prose at the post-secondary level: the influence of textual variables on L2 reading comprehension (a genre-based approach). *Reading in a Foreign Language* 8, 645–62.

Samson, D.M.M. 1983: Rasch and reading. In van Weeren, H., editor, *Practice and problems in language testing*. Arnhem: CITO.

Sarig, G. 1989: Testing meaning construction: can we do it fairly? *Language Testing* 6, 77–94.

Shohamy, E. 1984: Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1, 147–70.

Shohamy, E. and **Inbar, O.** 1991: Validation of listening comprehension tests: the effect of text and question type. *Language Testing* 8, 23–40.

Steffensen, M.S. and **Joag-Dev, C.** 1984: Cultural knowledge and reading. In Alderson, J.C. and Urquhart, A.H., editors, 48–61.

Steffensen, M.S., Joag-Dev, C. and **Anderson, R.C.** 1979: A cross-cultural perspective on reading comprehension. *Reading Research Quarterly* 15, 10–29.

Taylor, B.M. and **Samuels, S.J.** 1983: Children's use of text structure in the recall of expository material. *American Educational Research Journal* 20, 517–28.

Trabasso, T., Secco, T. and **van den Broek, P.W.** 1984: Causal cohesion and story coherence. In Mandl, H. *et al*., editors, *Learning and understanding from text.* Hillsdale, NJ: Lawrence Erlbaum, 83–111.

Ulijn, J.M. and **Strother, J.B.** 1990: The effect of syntactic simplification on reading EST texts as L1 and L2. *Journal of Research in Reading* 13, 38–54.

Urquhart, A.H. 1984: The effect of rhetorical ordering on readability. In Alderson, J.C. and Urquhart, A.H., editors, 160–75.

Weir, C.J. 1993: *Understanding and developing language tests*. London: Prentice Hall.

Zuck, L.V. and **Zuck, J.G.** 1984: The main idea: specialist and non-specialist judgments. In Pugh, A.K. and Ulijn, J.M., editors, *Reading for professional purposes*. London: Heinemann, 130–35.

Appendix 1 Sample test

(Words in brackets are the words deleted for a cloze test.)

The industrialised countries between them possess 78% of all existing wealth in the world. This means that the other countries, which are usually (called)[1] the 'Third World', have about 22%, even though their population is (about)[2] 76% of the world's total. Many rich industrialised countries give aid to poorer (Third)[3] World countries. The intention is simple – giving aid in this way should (help)[4] the poorer countries to improve their situation. Of course they hope that (aid)[5] will no longer be necessary in the end, since the Third World (countries)[6] will have become able to look after themselves.

However, many people argue (that)[7] much of the aid given to Third World countries does more harm (than)[8] good. One example of this is 'tied aid'. Money or machinery is (given)[9] to a Third World country, but on certain conditions. These usually mean, (for)[10] example, that the receiving country has to spend the money on what (is)[11] produced in the giving country. As a result, the Third World country (may)[12] have to buy products it does not need, or at a higher (price)[13].

At the same time Third World countries become dependent on industrialised countries. (They)[14] need them more and more. For example, a Third World country may (be)[15] given expensive tractors. Agricultural productivity may improve enormously, but when the tractors (go)[16] wrong, they will require skilled mechanics or expensive spare parts. Either way, (the)[17] poor country needs to pay money to the richer country to repair (the)[18] tractors.

Moreover, most aid has been used in cities. This makes life (there)[19] look more attractive, offering jobs which are highly paid and which are (not)[20] available in rural areas. So people leave the countryside and move to (cities)[21]. As a result, the countryside becomes empty and the country can no (longer)[22] produce enough food for its people. At the same time, cities become (overcrowded)[23] and there are all sorts of problems, from housing shortages to poor (health)[24] facilities. Worse still, there may not be enough jobs for all the (people)[25] who come to the cities hoping that they will become richer: many of them, in fact, become poorer than before.

Open-ended questions

Answer the following questions *in Japanese*.

1. When they give aid to Third World countries, what do industrialised countries want to happen in the future? (Literal understanding; Local)

2. What is 'tied aid'? (Integration; across sentences)
3. Why do Third World countries need to pay more money to richer countries when tractors go wrong? (Literal understanding/ integration; across sentences)
4. Why is it a problem if people move from the countryside to cities? (Integration; across sentences)
5. Does the writer of the passage think that giving aid is generally successful? Give reasons. (Inference; Global)

Summary writing

Write a summary of the passage in about 100 Japanese characters.

Appendix 2

Descriptive statistics of the reading comprehension tests (converted in percentages)

	$\bar{x}$ (s.d.)	$\bar{x}$ (s.d.)	$\bar{x}$ (s.d.)	Range of facility values
Cloze ($N = 255$)	*Aid* ($n = 25$)	*Sea safety* ($n = 25$)	*Total* ($n = 50$)	
Association ($N = 63$)	46.0 (15.1)	38.7 (17.8)	42.3 (15.3)	.02–.95
Description ($N = 66$)	42.7 (18.7)	36.9 (16.7)	39.5 (17.0)	.02–.86
Causation ($N = 61$)	37.4 (19.9)	39.7 (17.7)	38.6 (17.9)	.00–.97
Problem–solution ($N = 65$)	34.2 (17.3)	29.4 (15.0)	31.8 (15.3)	.02–.89
Open-ended ($N = 227$)	*Aid* ($n = 5$)	*Sea safety* ($n = 5$)	*Total* ($n = 10$)	
Association ($N = 59$)	32.6 (18.0)	34.1 (22.5)	33.4 (17.9)	.08–.66
Description ($N = 54$)	48.8 (23.9)	49.6 (21.9)	49.1 (20.1)	.29–.75
Causation ($N = 57$)	50.6 (25.5)	43.2 (27.4)	46.9 (23.3)	.19–.68
Problem–solution ($N = 57$)	46.3 (21.9)	43.6 (27.3)	44.9 (22.6)	.15–.67
Summary ($N = 253$)	*Aid* ($n = 10$)	*Sea safety* ($n = 10$)	*Total* ($n = 20$)	
Association ($N = 66$)	29.2 (17.5)	32.7 (20.5)	30.9 (15.7)	.00–.62
Description ($N = 62$)	38.5 (20.9)	28.6 (18.7)	33.6 (16.4)	.02–.60
Causation ($N = 63$)	40.2 (20.8)	43.5 (22.5)	41.8 (19.3)	.02–.87
Problem–solution ($N = 62$)	36.6 (19.4)	40.0 (19.8)	38.4 (16.4)	.00–.69

Notes: N = number of participants; n = number of items.

Item total correlations: range and mean

	Range	Mean
Cloze (*N* = 255)	(*n* = 50)	
Association (*N* = 63)	.00–.53	.29
Description (*N* = 66)	.07–.59	.33
Causation (*N* = 61)	.06–.62	.35
Problem–solution (*N* = 65)	.00–.53	.31
Open-ended (*N* = 227)	(*n* = 10)	
Association (*N* = 59)	.11–.62	.35
Description (*N* = 54)	.14–.58	.40
Causation (*N* = 57)	.35–.65	.47
Problem–solution (*N* = 57)	.26–.68	.50
Summary (*N* = 253)	(*n* = 20)	
Association (*N* = 66)	.00–.63	.34
Description (*N* = 62)	.00–.73	.36
Causation (*N* = 63)	.00–.82	.41
Problem–solution (*N* = 62)	.00–.79	.36

Notes: *N* = number of participants; *n* = number of items.

Appendix 3 Reading comprehension test results of three language proficiency groups and the results of one-way analysis of variance

Cloze tests (*N* = 255)

	Low (*N* = 75) $\bar{x}$ *(s.d)*	Middle (*N* = 86) $\bar{x}$ *(s.d.)*	High (*N* = 94) $\bar{x}$ *(s.d.)*	ANOVA *F (d.f.)*
Association (*N* = 63)	29 (11)	40 (12)	54 (13)	21.16** (2, 60)
Description (*N* = 66)	20 (13)	42 (12)	53 (8)	44.59** (2, 63)
Causation (*N* = 61)	26 (14)	37 (13)	53 (15)	19.96** (2, 58)
Problem–solution (*N* = 65)	16 (9)	32 (12)	43 (10)	36.12** (2, 62)

Notes: $^{**}p < .001$.

Open-ended questions ($N = 227$)

	Low ($N = 82$) $\bar{x}$ (s.d)	Middle ($N = 71$) $\bar{x}$ (s.d.)	High ($N = 74$) $\bar{x}$ (s.d.)	ANOVA F (d.f.)
Association ($N = 59$)	23 (11)	37 (19)	43 (19)	7.87* (2, 56)
Description ($N = 54$)	36 (18)	52 (14)	65 (19)	11.53** (2, 51)
Causation ($N = 57$)	25 (19)	48 (16)	65 (16)	26.11** (2, 54)
Problem–solution ($N = 57$)	25 (16)	49 (18)	61 (14)	32.76** (2, 54)

Notes: $^{**}p < .001$; $^{*}p < .005$.

Summary writing ($N = 253$)

	Low ($N = 82$) $\bar{x}$ (s.d)	Middle ($N = 81$) $\bar{x}$ (s.d.)	High ($N = 90$) $\bar{x}$ (s.d.)	ANOVA F (d.f.)
Association ($N = 66$)	21 (16)	35 (13)	36 (13)	7.07* (2, 63)
Description ($N = 62$)	23 (12)	36 (10)	43 (17)	12.49** (2, 59)
Causation ($N = 63$)	25 (17)	45 (10)	56 (17)	22.41** (2, 60)
Problem–solution ($N = 62$)	23 (13)	36 (11)	53 (11)	31.46** (2, 59)

Notes: $^{**}p < .001$; $^{*}p < .005$.