This study examined the effects of item characteristics on cloze test performance. A particular focus of the study was to investigate the relationship between a number of cloze item characteristics and scoring methods. The item characteristics examined were: (a) content/function words, (b) parts of speech, (c) word frequency, (d) the number of occurrences of a word in the text, (e) alternative answers, (f) syntactic variation, (g) the amount of context, and (h) knowledge base. A total of 255 Japanese university students participated and were randomly assigned 1 of 4 sets of cloze tests. The data revealed variations in the results of cloze tests attributable to different scoring methods. The study showed the complex effects of item characteristics and scoring methods on test performance, both in terms of item difficulty and discrimination. These findings have important implications for the use of cloze tests.

THE CLOZE PROCEDURE IS WIDELY USED IN tests of language ability. The procedure requires the test taker to fill in the blanks in a text in which a number of words have been deleted, usually at specific intervals. The deletion of words at regular intervals ostensibly produces a representative sample of the linguistic features of the text, which makes it possible to obtain a valid measure of the test-taker’s underlying language ability.

Past research has focused on exactly which aspects of language ability are measured by the test, and different studies have generated widely varying results. Many of these studies have examined the effect of the text (e.g., whether the difficulty of the text has an effect on the performance), the starting point for deletion, the scoring method, and the deletion rate (Alderson, 1978, 1979a, 1979b, 1980, 1983; Porter, 1978). For example, changing the deletion rate seems to result in significantly different mean scores for some texts with some scoring procedures, but not with other texts or other procedures (Alderson, 1980). This variance in deletion rate suggests that not all cloze tests measure exactly the same language abilities, and as Alderson (1979a) says, “individual cloze tests vary greatly as measures of EFL proficiency” (p. 225). However, these studies did not go so far as to examine what might have caused this variation.

Some other problems with cloze tests are: (a) They seem to measure “lower-order” language skills rather than “higher-order” skills (Alderson, 1979b; Chihara, Oller, Weaver & Chavez-Oller, 1977); (b) even native speakers may not be perfectly correct, especially when the exact-word scoring method is used (Alderson, 1980; Oller, 1979); (c) understanding a text and being able to fill in blanks are not the same thing, the latter requiring more productive skill; and (d) factors other than language ability seem to contribute to test performance (Chapelle & Abraham, 1990; Stansfield & Hansen, 1983; Turner, 1989).

PREVIOUS STUDIES ON CLOZE TESTS

Along with the more general cloze research, there have also been attempts to identify the nature of cloze items so that we can see more clearly what each item measures. There are two trends...
Within this area, one is to try to establish consistency across different versions of cloze tests by looking at certain item characteristics (Bachman, 1982, 1985; Jonz, 1990; Lee, 1985). The other is to try to characterize the relationship between item characteristics and item difficulties (Abraham & Chapelle, 1992; Brown, 1989, 1993).

Two studies by Bachman (1982, 1985) revealed no significant differences among different versions in the relative frequency of the four categories of item types he devised. Jonz (1990) found a similar result in his earlier analysis of eight fixed-ratio cloze tests, which suggested that the cloze procedure was “far from erratic in its selection of item types” (p. 72).

These findings are noteworthy because they support claims of consistency across different fixed-ratio cloze tests in terms of item characteristics. If this finding is confirmed, we will be able to use cloze tests with more confidence. However, problems remain. One, as Jonz (1990) himself accepts, is the classification framework by which he identifies item types. The framework he developed on the basis of Bachman’s earlier study (1985) did not cover the whole range of cloze items and needs further development. Furthermore, we have yet to establish a reliable and objective way to categorize cloze items in terms of their characteristics. Jonz’s categorization was based solely on his own judgment, and he mentions in passing discrepancies between his own categorization and those of Bachman and other analysts. Clearly, his study needs to be extended in several ways (e.g., more people need to be involved in the judgment of item categorization). The present study addressed this issue by characterizing items on the basis of agreement among experts.

Furthermore, the main purpose of both Bachman’s and Jonz’s studies was to establish the comparability of different cloze tests. In line with this objective, their frameworks cover no more than a limited range of cloze item characteristics. For example, they do not include the familiarity of the word to be filled in. When cloze tests are used with second language (L2) learners, there is a significant likelihood that some words will be outside learners’ vocabulary and that these items will inevitably be extremely difficult, regardless of the amount of context required. Frequency is another important issue: Even if the word to be filled in is an unfamiliar one, it will not be so difficult to restore it if it is closely related to the main topic of the passage, or if the same word is repeated many times in the passage. These characteristics must definitely influence item difficulty. However, only a few studies have examined their influence.

Among these few studies are those by Brown (1989, 1993) and Abraham and Chapelle (1992). Brown examined the relationship between cloze item characteristics and item difficulty. He reported that the item difficulty was closely related to the frequency of the word in the passage, word length, the number of occurrences of a test item word in the passage, the text difficulty, the length of a word to be restored, the position of the deletion in the text, and the number of syllables per T-unit.1

Abraham and Chapelle (1992) examined item characteristics in some detail, as well as the effects of different cloze formats: fixed-ratio, rational,2 and multiple-choice cloze. The item characteristics they examined included: (a) whether the deleted word is a content or function word, (b) the word length in terms of the number of letters, (c) whether alternative answers were possible, (d) the number of occurrences of the word in the text, (e) the inflectional morpheme, and (f) the amount of context required to restore the word. The study detected some relationships between item characteristics and item difficulty. For example, the amount of context and the length of words correlated negatively with item difficulty in the rational deletion cloze test. This finding suggests that items tend to be easier when a smaller amount of context is necessary to restore the words and when the length of the words is shorter. Abraham and Chapelle also found that function words were easier than content words both in the fixed-ratio and rational deletion cloze, but not in the multiple-choice cloze.

This kind of research is extremely valuable because it enables us to identify the relationship between particular cloze item characteristics and item difficulty. It would, therefore, be useful to explore this issue further and learn more about the effects of cloze item characteristics on test performance. The research studies conducted so far have only examined the relationship between item characteristics and item difficulty: it would be more illuminating to examine what kinds of cloze item reflect the learners’ language ability more accurately than others and, thus, help to establish their validity. The present study sought to meet this objective in its inclusion of discrimination indices in the analysis.

PURPOSE OF THE STUDY

The purpose of the present study was to examine the effects of item characteristics on cloze test performance with particular reference to reading comprehension tests.3 In particular, the study fo-
cused on important characteristics of cloze items, with special attention on the scoring methods. To date, few studies have investigated cloze tests with regard to the effects of scoring methods. Brown’s study (1980) is probably the most extensive and systematic. His findings suggest that the acceptable-word scoring method is more reliable and valid. However, he was concerned with the overall results of cloze tests, not with item-level test performance. The present study further explored this area and examined whether and how different scoring methods would affect cloze items in respect to item characteristics. Drawing on the earlier work of Abraham and Chapelle (1992), these characteristics were: (a) content and function words, (b) parts of speech, (c) word frequency, (d) the number of occurrences of a word in the text, (e) alternative answers, (f) syntactic variation, (g) the amount of context, and (h) knowledge base.

A pilot study showed that the content/function word distinction was by far the most important factor influencing test performance. I, therefore, examined this distinction first, and cross-referenced it with the analysis of other categories.

In addition to the content/function word distinction, item words were more closely examined in terms of different parts of speech, because this more traditional classification might reveal further distinctions among words. Even among function words, some parts of speech seem to have more meanings than others or more forms within a limited set. This variation could affect cloze item difficulty or efficiency in discrimination.

The next three categories for analysis were: word frequency, the number of occurrences, and alternatives. The Cambridge English Lexicon (Hindmarsh, 1980) served as a measure of word frequency. The Lexicon was compiled on the basis of a number of word frequency lists including West (1953) and Thorndike and Lorge (1944), listing words from Level 1 (most frequent) to Level 7 (least frequent). The number of occurrences are categorized into three levels: (a) The same word does not appear elsewhere, (b) the same word appears from two to four times, and (c) the word appears more than five times in the text. The category “Alternatives” indicates whether alternative answers are possible or not. For example, in the case of content words (e.g., verbs, nouns, and adjectives) alternative words are often possible in the given context and acceptable both semantically and syntactically.

Syntactic variation was included to allow for further comparison, especially in relation to the scoring methods. The purpose was to permit examination of word classes like nouns and verbs that show syntactic variation, such as plural forms or verb endings.

For the last two categories, the amount of context and knowledge base, five-level classification frameworks based on Bachman (1985), Jonz (1991), and Abraham and Chapelle (1992) were used. Details of the five-level classification are shown in the sections that describe the results.

To enhance the reliability of the classification, six experts were invited to make judgments concerning the last two categories. The degree of agreement in their judgments varied as a function of the items: Some items achieved unanimous agreement, and others generated complete disagreement. As Jonz (1991) and Abraham and Chapelle (1992) indicated, it is not a straightforward matter to decide how much context or what kind of knowledge is required to complete a gap. In this study, instead of forcing every item into a particular type, a more flexible method was adopted: Items were categorized as requiring a certain amount of context or being related to a specific type of knowledge when they met the criterion of agreement among at least 4 of 6 judges (66.7% or higher). Other items were left as unidentifiable. This approach permitted a focus on those items for which consensus was possible, while setting aside, for the moment, those items associated with a more confused picture.

RESEARCH METHOD

Piloting

Before the main data collection, a series of preliminary studies was conducted to establish the feasibility of the study and to identify potential pitfalls in the proposed methodology. To this end, the influence of a number of relevant variables was explored. These variables included: topic areas of reading passages, text length, text readability, students’ language proficiency, deletion ratio, the nature of words to be restored (especially the ratio between content words and function words), and the number of occurrences of content words. The largest preliminary study involved 219 Japanese university students, and a number of smaller-scale studies involved judgments by selected English language teaching experts.

Participants

A total of 255 Japanese university students, 91 male and 164 female, participated in the main study. They were all native speakers of Japanese,
attending one of eight universities in Japan. The majority were 18–19 years of age and in the first or second year of their courses. They had completed 6 years of English instruction at secondary school, and they were fairly homogeneous in their social and educational backgrounds.

**Materials**

Eight texts were selected for the purpose of a larger study involving examination of the effects of rhetorical organization of texts on performance in reading comprehension tests (Kobayashi, 1995, 2002).

The results of the pilot study helped identify a number of factors, such as the appropriate text length and difficulty level for the participants. The topic areas were chosen so that they would be neither too familiar nor too unfamiliar to the participants. This decision was made in order to avoid the undesirable effects of background knowledge. Out of several possible passages, two texts concerning international aid and marine safety were selected from educational materials. Each of these two texts served as a source from which to provide four variant texts with different kinds of rhetorical organization. Special care was taken to ensure that each of the texts was standardized with respect to length and difficulty level. The average length of the four texts was 357.8 words for the international aid texts and 380.5 words for the marine safety texts. The average Flesch-Kincaid readability indices were 8.4 and 8.2, respectively.

Further decisions made on the basis of the pilot results were:

1. Fixed-ratio cloze tests were prepared, with a deletion rate of every 13th word. This procedure resulted in 25 blanks per text. Even though more than 25 blanks were possible with longer texts, the number of items was made consistent across texts to make the later analysis easier. All together, there were 200 items to be examined (25 items by 8 texts).

2. The starting point for deletion varied from one text to another, instead of starting at a particular point in the text, as is often the case, so that each of the texts would be as similar as possible with respect to the ratio of function and content words, and also with regard to the number of occurrences of content words in the text. This decision followed an extensive examination of the texts that revealed that changing the starting point for deletion resulted in markedly different sets of words to be deleted.

3. It was decided to avoid deleting proper nouns and numbers. When a deletion fell on a proper noun or a number, the preceding word was chosen for deletion. If the preceding word was also a proper noun or a number, then the subsequent word was chosen instead. A sample test is shown in Appendix A.

In addition to the cloze tests, participants completed a 50-item, multiple-choice English language proficiency test. The items were selected and modified from the Cambridge First Certificate Examination practice books (e.g., Archer & Nolan-Woods, 1983) and a proficiency test for overseas students administered at a British university. The purposes of this test were to ensure the comparability of the four groups of participants in their general English language ability and to categorize the participants into three different proficiency groups according to the rank order of their scores as a basis for comparison at a later stage of analysis. The relevant statistical data were: Cronbach’s alpha ($\alpha = .82$; facility values ranging from .17 to .99, with a mean of .59; and item-total correlations ranging from .08 to .53, with a mean of .34.

**Procedure**

Data collection took place in intact English-language classes under the supervision of classroom teachers who were given detailed written directions. Written instructions were also prepared for students. Both sets of instructions were in Japanese and piloted to minimize misunderstanding or confusion.

The participants were randomly divided into four groups; the four sets of two cloze passages were arranged so that each test would be randomly distributed among the students. This procedure ensured that no students read a similar text more than once. Otherwise, the validity of the research might have been undermined. Statistical analysis based on the proficiency test results showed no significant differences among the four groups in terms of their general English ability ($F = 0.09$, df = 3, 251, n.s.).

To control for any possible order effect, the order of the two texts in each set was counterbalanced. In the light of the pilot results, 50 minutes were allowed for completing the two cloze tests, making a total of 50 cloze items.

Apart from the main data collection regarding students’ performances for cloze tests, an examination of individual cloze items was conducted in consultation with six experts to ensure the reli-
ability of item categorization (cf. Jonz, 1990). All the judges used the eight cloze tests and a five-level framework to categorize cloze items in terms of the amount of context and kind of knowledge required to complete the gaps. (See the result sections for the details of the five levels.)

Analysis

All the papers were marked twice by the researcher, with an educated native speaker’s assistance. Following the usual convention, minor orthographical mistakes were ignored. To permit the examination of the interaction between the scoring methods and item characteristics, three different scoring methods were employed: (a) the exact word scoring method (Ex); (b) the semantically and syntactically acceptable word scoring method (Acc1); and (c) the semantically acceptable, but syntactically unacceptable word scoring method (Acc2). Acceptable alternative answers had been collected in consultation with six educated native speakers of English prior to the scoring. (See Appendix B for sample alternative answers.)

The data were analyzed using SPSS/PC. The analysis included: calculation of means, standard deviations, reliability for the complete set of items, facility values and item-total correlations (used as a means of establishing discrimination indices) for individual items, and correlations with the proficiency test. Statistical analysis also included t-tests and correlational analyses when these were deemed useful.

RESULTS

Comparison of Three Scoring Methods

All answers were scored according to the three different scoring methods mentioned above, and descriptive statistics were calculated for each scoring method (Appendix C). This section offers an exploratory comparison of the overall results of the cloze tests by the three scoring methods. The following section, which covers the main part of the research, provides a detailed analysis of cloze item characteristics in relation to the three scoring methods.

Not surprisingly, the results showed that the acceptable-word scoring methods (Acc1 and Acc2) led to higher mean scores. The semantically acceptable-word scoring method yielded slightly, but consistently, higher scores.

The two acceptable-word scoring methods had consistently higher reliability, although the differences were not very great. It is particularly interesting to examine the reliability of the content and function word items separately (Appendix D). On the whole, content word items had lower reliability estimates when the exact-word scoring method was applied. However, values rose dramatically when acceptable-word scoring methods were applied, and the values were even higher than those of function word items in many of the texts. Of course, further research is needed to confirm this pattern, but this finding seems to suggest the possibility that acceptable-word scoring methods are particularly effective in capturing useful variance for content word items. (See the discussion below for further comparison between content and function words.)

The correlations (r) with the proficiency test were moderately high overall. However, the correlations were slightly higher for the acceptable-word scoring methods, especially the semantically-only acceptable-word scoring method. This finding suggests convergent validity for the cloze tests, and indicates some overlap in what the cloze and proficiency tests measure, especially when acceptable-word scoring methods are adopted.

The correlations among the different scoring methods were consistently high, with only 1 of 24 correlations below .90 (at .88—Appendix E). Correlations between the exact-word method and semantically acceptable-word method were slightly lower than others.

A Comparison across Proficiency Levels

The most interesting finding concerns the differences in mean scores. Figure 1 shows the variations in the scores (in percentages) for the three proficiency groups according to the different scoring methods.

There was a slight tendency for higher groups to gain more in the acceptable-word scoring methods. Gains in scores of each group from the exact-word scoring method to the acceptable-word scoring method were calculated in both raw scores and percentages against the original scores and shown in Table 1.

The table shows that the gains from the exact-word scoring method (Ex) to the acceptable-word methods (Acc1 and Acc2) were greater for higher groups. This finding suggests that students with higher proficiency may be unfairly penalized when the exact-word scoring method is applied. This result is probably because those with greater language proficiency are more able to think of suitable words that are not the original
ones, but nonetheless make sense in the context, as native speakers might. However, if the exact-word scoring method is applied, some of their answers become incorrect solely because they are not the precise originals (Alderson, 1980; Oller, 1979).

Table 1 also shows another interesting trend: Lower-proficiency groups benefited more when the semantically only acceptable-word scoring method (Acc2) was applied than the higher group. Even though the actual gain in scores was not great, the gain in percentages was greatest in the lower group (8.4% from Acc1 to Acc2). This finding suggests that lower-proficiency learners may have had more problems at a syntactic level than higher-proficiency learners. It is interesting that the middle group benefited from both of the acceptable-word methods. This benefit may have occurred because this group of learners still had syntactic problems but were closer to the higher group in that they could think of alternative answers that could make sense in the context.

The findings presented here seem to support the view that the acceptable-word scoring method is fairer, especially to those with greater language proficiency (Alderson, 1980; Brown, 1980). The differences among the proficiency groups observed here may not be striking, but it would certainly be interesting to investigate further how the difference changes as more proficient learners or native speakers are included in the comparison.

A CLOZE ITEM ANALYSIS IN TERMS OF ITEM CHARACTERISTICS

This section presents the main focus of the study: close investigation of cloze items in terms of eight item characteristics derived from previous studies (e.g., Abraham & Chapelle, 1992) and their relationships to different scoring methods.

The following pages examine facility values and item-total correlations (used as a means of establishing discrimination indices) of cloze items, by the three scoring methods. FV1, FV2, and FV3 each represent facility values by the exact-word scoring method, the semantically and syntactically acceptable-word scoring method, and the semantically only acceptable-word scoring method, respectively. The same is true of IT1, IT2, and IT3 in respect of item-total correlations. Due to space limitations, not all the relevant statistics are included here (see Kobayashi, 1995, for details).

### TABLE 1
Gains in Scores and Percentages of the Three Proficiency Groups

<table>
<thead>
<tr>
<th></th>
<th>From Ex to Acc1</th>
<th>From Ex to Acc2</th>
<th>From Acc1 to Acc2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher</td>
<td>13.1 (35.2%)</td>
<td>15.3 (41.7%)</td>
<td>2.2 (4.4%)</td>
</tr>
<tr>
<td>Middle</td>
<td>9.9 (35.0%)</td>
<td>12.2 (43.1%)</td>
<td>2.3 (6.0%)</td>
</tr>
<tr>
<td>Lower</td>
<td>5.2 (29.9%)</td>
<td>7.1 (40.8%)</td>
<td>1.9 (8.4%)</td>
</tr>
</tbody>
</table>

Note. Ex = Exact-word scoring method; Acc1 = Semantically and syntactically acceptable-word scoring method; Acc2 = Semantically acceptable, but syntactically unacceptable-word scoring method.
Content versus Function Words

As shown in Table 2 below, function words were easier, regardless of scoring methods. The difference between content words and function words (calculated as the means of their facility values) was statistically significant: \( t = 3.66, 2.81, 2.43, df = 192, p < .01 \) for FV1, FV2, and FV3, respectively. This result replicates the findings of Abraham and Chapelle’s 1992 study and is understandable because learners tend to learn function words at an earlier stage of their language development.

However, the rate of increase in facility values is greater for content words when scoring methods changed from the exact-word method to the acceptable-word method (50% vs. 23% increase in content vs. function word items, respectively). The variance produced by these changes in scoring methods implicts for reliability, given that the item-total correlations also rose from IT1 to IT3 by .06 for content words and .05 for function words.

Parts of Speech

Relative pronouns, pronouns, and articles proved to be the most difficult types of function words. Articles present particular difficulty, possibly because Japanese learners do not have articles in their mother tongue and, therefore, do not realize that articles are necessary when they are missing.

Content words (i.e., nouns, adjectives, adverbs, and verbs) showed a similar tendency: Their facility values varied markedly according to the scoring methods (e.g., FVs of nouns: .25, .34, .35 for Ex, Acc1, and Acc2, respectively). In particular, nouns and verbs became much easier when syntactic variation was accepted as correct. This observation seems to be related to the issue of meaning versus syntactic accuracy. Learners probably do not prioritize syntactic variation when they are more concerned with meanings. (See Sato, 1985, for the effects of task variations on L2 learners’ performance; also see Skehan, 1998, for L2 learners’ dual coding.) When the semantically only acceptable-word method was applied (FV3), almost all content words came in the middle range of an overall ranking of difficulty, with two distinct groups of function words at either end of the difficulty continuum: Articles and relative pronouns proved to be the most difficult, and prepositions and to-infinitives the easiest.

Item-total correlations were higher for word-classes such as to-infinitives, adverbs, and conjunctions. This finding may be explained by the fact that these words are related to some kind of textual organization requiring a high level of language ability. However, item-total correlations for relative pronouns and pronouns, which are also supposed to be related to textual organization, were low. This result is hard to explain, but may have occurred because these words are more closely related to a local level of cohesion rather than to a more global organization. However, the number of cases involving relative pronouns was very small \( (n = 3) \); therefore, it may not be possible to draw any general conclusion.

Articles showed by far the lowest item-total cor-

<table>
<thead>
<tr>
<th>Parts of Speech</th>
<th>( N ) (%)</th>
<th>FV1</th>
<th>FV2</th>
<th>FV3</th>
<th>IT1</th>
<th>IT2</th>
<th>IT3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>91a (45.5)</td>
<td>.22</td>
<td>.33</td>
<td>.36</td>
<td>.27</td>
<td>.32</td>
<td>.33</td>
</tr>
<tr>
<td>Function</td>
<td>103 (51.5)</td>
<td>.35</td>
<td>.43</td>
<td>.44</td>
<td>.28</td>
<td>.32</td>
<td>.33</td>
</tr>
<tr>
<td>Mixedb</td>
<td>6 (3.0)</td>
<td>.06</td>
<td>.41</td>
<td>.42</td>
<td>.14</td>
<td>.28</td>
<td>.29</td>
</tr>
<tr>
<td>TOTAL</td>
<td>200 (100)</td>
<td>.28</td>
<td>.38</td>
<td>.40</td>
<td>.27</td>
<td>.32</td>
<td>.33</td>
</tr>
</tbody>
</table>

Note: FV1, FV2, and FV3 represent facility values by the exact-word scoring method, the semantically and syntactically acceptable-word scoring method, and the semantically only acceptable-word scoring method, respectively. The same is true of IT1, IT2, and IT3.

aEven though every effort was made to make the ratio of content and function words equal in the preparation of the study, there were, in fact, slightly more function words than content words. Originally there were 12 content-word items and 13 function-word items. However, in two content-word items, no one restored the original words but instead gave function words that were both grammatically and semantically acceptable in the context. Therefore the two items were re-categorized as function-word items, and the balance between content and function words was disrupted in this text.

bIt turned out that six items could be filled in by either content or function words when the acceptable-word scoring method was adopted. These items were classified as ‘mixed’ and exempted from the subsequent comparative analysis.
relations. As noted above, this low correlation seems to be related to the particular difficulties faced by Japanese learners of English. Even learners with high proficiency have problems with articles and thus cloze items related to articles may not be a reliable measure of language proficiency. This area is potentially interesting for further research with learners of different first languages.

The item-total correlations for adjectives, verbs, and prepositions were fairly low. In the case of prepositions, one possible explanation may be that they are often recoverable by local understanding, such as the knowledge of collocation or set-phrases, which may not reflect learners’ overall language proficiency.

**Word Frequency**

Generally there was a tendency for more frequent words to be easier to restore, especially when the exact-word scoring was applied ($r = -0.19, p < .01$), but the relationship was not significant in other scoring methods. This result is not surprising because, in the acceptable-word method, learners are allowed to fill the blanks with any word that makes sense in the context, and these words may not reflect the frequency level of the original words.

The majority (62.1%) of function words were at level 1 (most frequent in the Cambridge English Lexicon), and no function words were beyond level 4. There did not seem to be a frequency effect with such words. Item-total correlations were slightly higher at frequency levels 2 and 3.

**Occurrences**

The results (Table 3) show that the more frequently the same words appeared elsewhere in the text, the easier the items became, especially when the exact-word scoring method was applied. This observation applies particularly to content words and may be explained by Finn’s (1978, p. 508) information theory which holds that “rare words that are repeated in a text are closely associated with the topic of the text.” This finding is supported statistically ($r$ [correlations between the number of occurrences and item difficulty] = .41, .23, .23, $p < .001$ for FV1, FV2, and FV3, respectively).

Nearly half (48.4%) of the content words appeared only once, and such items were difficult, explaining, in part, perhaps the lower facility values of content-word items. On the other hand, facility values of content words that appeared five times or more were far higher than those of other content words. The difference in the three levels of occurrence was particularly remarkable in the exact-word scoring method, perhaps because the frequent occurrence of the words enables learners to identify them and give the exact words as answers. However, when the acceptable-word scoring method was adopted, content words that appeared fewer than five times became much easier, especially those that appeared only once. It is more likely that, when the words do not appear elsewhere in the text, learners may not be able to think of the exact words but choose other semantically acceptable words instead.

| TABLE 3 |
| Facility Values and Item-Total Correlations for the Number of Occurrences of the Word in the Text |

<table>
<thead>
<tr>
<th>Total $(n = 200)$</th>
<th>Occurrence</th>
<th>N</th>
<th>%</th>
<th>FV1</th>
<th>FV2</th>
<th>FV3</th>
<th>IT1</th>
<th>IT2</th>
<th>IT3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59</td>
<td>29.5</td>
<td>.15</td>
<td>.32</td>
<td>.34</td>
<td>.25</td>
<td>.31</td>
<td>.31</td>
<td></td>
</tr>
<tr>
<td>2–4</td>
<td>71</td>
<td>35.5</td>
<td>.26</td>
<td>.36</td>
<td>.37</td>
<td>.27</td>
<td>.34</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>5 +</td>
<td>70</td>
<td>35.0</td>
<td>.41</td>
<td>.46</td>
<td>.48</td>
<td>.28</td>
<td>.31</td>
<td>.32</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content Words $(n = 91)$</th>
<th>Occurrence</th>
<th>N</th>
<th>%</th>
<th>FV1</th>
<th>FV2</th>
<th>FV3</th>
<th>IT1</th>
<th>IT2</th>
<th>IT3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44</td>
<td>48.4</td>
<td>.14</td>
<td>.30</td>
<td>.32</td>
<td>.27</td>
<td>.32</td>
<td>.32</td>
<td></td>
</tr>
<tr>
<td>2–4</td>
<td>34</td>
<td>37.4</td>
<td>.24</td>
<td>.31</td>
<td>.32</td>
<td>.26</td>
<td>.33</td>
<td>.33</td>
<td></td>
</tr>
<tr>
<td>5 +</td>
<td>13</td>
<td>14.3</td>
<td>.50</td>
<td>.50</td>
<td>.57</td>
<td>.28</td>
<td>.31</td>
<td>.32</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function Words $(n = 103)$</th>
<th>Occurrence</th>
<th>N</th>
<th>%</th>
<th>FV1</th>
<th>FV2</th>
<th>FV3</th>
<th>IT1</th>
<th>IT2</th>
<th>IT3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>11.7</td>
<td>.22</td>
<td>.38</td>
<td>.40</td>
<td>.23</td>
<td>.29</td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>2–4</td>
<td>35</td>
<td>34.0</td>
<td>.30</td>
<td>.41</td>
<td>.42</td>
<td>.29</td>
<td>.36</td>
<td>.36</td>
<td></td>
</tr>
<tr>
<td>5 +</td>
<td>56</td>
<td>54.4</td>
<td>.41</td>
<td>.45</td>
<td>.46</td>
<td>.28</td>
<td>.31</td>
<td>.32</td>
<td></td>
</tr>
</tbody>
</table>
The majority (54.4%) of the function words appeared five times or more elsewhere in the text. This result was not surprising considering the nature of function words. When the exact-word method was used, facility values were higher for words that appeared many times than for words that did not, as in the case of content words. However, when the acceptable-word method was applied, the effect of the number of occurrences was no longer clear or became almost nonexistent.

**Alternative Answers**

Items that allowed alternative answers were more difficult than those that did not, regardless of the scoring methods (e.g., $FV1 = .21$ vs. $.44; FV2 = .36$ vs. $.44$, respectively). This finding contradicts those of Abraham and Chapelle (1992). In their study, items that allowed alternative answers proved to be easier than those that did not. They argued that “items in which a single answer will fit are predicted to involve greater difficulty in retrieval than items which have more than one appropriate answer” (p. 470). However, items that allow alternative answers may require all the more cognitive ability because of the wide range of possible answers. Furthermore, such items are often content words, which, as noted above, are generally more difficult than function words. One explanation for the discrepancy may be the difference in the scoring methods employed: Abraham and Chapelle used the acceptable-word scoring method only. The data in the present study revealed that the difference between the facility values of items that allowed alternatives and those that did not was greatest when the exact-word scoring method was used. On the other hand, in the acceptable-word scoring methods, the difference between the two became progressively smaller and no longer statistically significant ($t = 6.19$, n.s., $n.s., p < .01$, in FV1, FV2, and FV3, respectively).

Item-total correlations were low in the exact-word method, especially when alternative answers were possible. This finding further supports the claim that it is fairer to accept words that make sense in a given context than to accept only exact-word matches.

As for items that do not have potential alternative answers, there appears to be little or no difference across scoring methods, either in facility values or item-total correlations.

**Syntactic Variation**

Syntactic variation includes verb endings, tenses, and plural forms. The issue of interest is how test performance is affected by the way such variation is handled. The result showed an interesting pattern. On the one hand, when different syntactic forms were not accepted as correct, facility values for these items were lower than those for the items without syntactic variation ($FV1 = .30$ vs. $.25; FV2 = .39$ vs. $.35$, respectively). On the other hand, when syntactic variation was considered acceptable, these items were easier than those for which such variation was not possible ($FV3 = .39$ vs. $.43$, respectively). The increase from FV1 to FV3 was considerable for items having syntactic variation—.18 (a 72% increase) in those items as opposed to .09 (a 30% increase) in items without syntactic variation—and the pattern was similar for both content and function words. This finding suggests that test performance on such items varies a great deal, depending on the scoring methods.

It must be remembered that the issue is about answers that were semantically acceptable. These answers showed that the learners at least understood the meaning of the surrounding context, but lacked productive skills in English syntactic forms or were simply careless. If it is the learners’ comprehension that is in focus, it seems justifiable to accept answers that are syntactically incorrect, but make sense in a given context. Even though the difference between the facility values of items having syntactic variation and those without in each scoring method was not statistically significant ($t = 1.14, 1.07, -0.90$ in FV1, FV2, and FV3, respectively), the finding seems to have a significant implication for the use of cloze tests as reading comprehension tests.

Generally, item-total correlations did not vary much according to whether the words had syntactic variation or not. However, in the case of function words, when different syntactic forms were considered acceptable, item-total correlations were higher. This result suggests that cloze tests become more effective in discriminating between more and less able learners when the semantically only acceptable-word scoring method is employed, especially in the case of function words.

**Amount of Context**

For the amount of context, the categories that were used were: (a) within a clause; (b) across clauses, but within a sentence; (c) across sentences, but within a paragraph; (d) beyond a paragraph and within the text; and (e) outside the text. Items were labelled as “I” when there was little or no agreement among the 6 judges (see above).

Unfortunately, as many as 56% of the content
were not clearly classified according to how much context was required to restore the words. This difficulty in identification may be related to the nature of content words. Since content words have an important role in conveying meanings and are linked to various other parts of a text in a complicated manner, it may become particularly difficult to pin down the exact context to which those words were related. This contrasts markedly with function-word items, 83.5% of which were clearly identified, and 69.9% of which were perceived as requiring a sentence-level context or smaller. Because the number of identified items was small, it is difficult to generalize, but content-word items seem to require more context than function-word items.

The data suggest that the smaller the required context, the easier the items become. The amount of context and facility values were negatively correlated, regardless of the scoring methods ($r = -.28, -.25, -.26$, $p < .001$, in FV1, FV2, FV3, respectively). This finding replicated those of previous studies (Abraham & Chapelle, 1992; Bachman, 1985). This relationship between the amount of context and item difficulty was clearly the case with function-word items. However, the situation was not so clear for content-word items. Items that could be filled in with the help of information contained in a single clause were certainly easy, but items that required context beyond a sentence were equally easy (e.g., .46 vs. .46 in FV2). This lack of a clear relationship may be related to a feature of content words mentioned above. Content words may be part of the lexical structure of a text (Hoej, 1991) that extends across several sentences. In such cases, the strength of the lexical ties or the semantic importance of the words in the text would be a more important determinant of the item difficulty than the amount of context. In future research, it would be worthwhile to investigate the relationship between the amount of context and the number of occurrences of a word in the text.

There was a tendency for items that require less context to show higher item-total correlations, in both content- and function-word items, but this was not statistically significant in any case ($r = .13, .02, .01$ in IT1, IT2, and IT3, respectively).

**Knowledge Base**

The kinds of knowledge considered in this study were: (a) knowledge of set phrases, (b) syntactic knowledge, (c) semantic knowledge, (d) knowledge of discourse, and (e) general knowledge.

On the whole, content-word items were more difficult to classify than function words. Only 64.8% of all content-word items were clearly perceived to be related to a particular type of knowledge. The difficulty of classification may, again, originate from the complex nature of content words. By contrast, 79.6% of the function-word items were identified as being related to a particular type of knowledge, and only 20.4% were left uncategorized.

Unsurprisingly, as many as 42.9% of content-word items were identified as being related to semantic knowledge. The facility values for these items were lower (.27 in FV2). By contrast, words identified as being related to knowledge of discourse were easier (.51 in FV2). This result is also understandable because these words appeared more than once, even several times, to construct lexical links in the text. On the whole, items not clearly identified showed higher item-total correlations. However, because the nature of these items was not clear, it is difficult to explain why this is so. Further research is necessary.

The majority (58.3%) of function-word items were related to syntactic knowledge. These words were generally easier (.43 in FV2) than content words, and their item-total correlations were higher (.35 in IT2). Specifically, 11.7% of function word items fell into the “knowledge of set phrases” category and they were fairly easy (.60 in FV2). On the other hand, their item-total correlations were relatively low (.28 in IT2). What is worth noting is that items related to set phrases in content words were not as easy as in function words (.36 vs .60 in FV2). This finding suggests that set phrases, which are comparatively straightforward, become still easier when function elements are missing, as opposed to when meaning elements are missing. It is also interesting that function-word items related to cohesion were difficult (.28 in FV2). This is a remarkable contrast to content-word items related to cohesion (.51 in FV2). This difference may have occurred because function words themselves have to create the connection of ideas and, therefore, become difficult to restore when missing, whereas cohesion-related content words constitute lexical ties in the text and are often linked to important ideas (Jonz, 1987).

**DISCUSSION**

The preceding sections have examined cloze test results obtained in the study in some detail, with a focus on different scoring methods and item characteristics, and have presented many interesting findings. Among other findings, three...
matters are worth emphasizing here: learners’ first language, the importance of meanings, and the framework for categorizing items.

In this study, definite and indefinite articles were a source of particular difficulty, and the item-total correlations for these items were also low. This observation seems to be closely related to the learners’ first language. Articles, whether definite or indefinite, do not exist in Japanese, and constitute an area in which Japanese learners always have difficulty when learning English. Learners, even those at a higher level of proficiency, have a tendency to omit articles without realizing that anything is missing. In such circumstances, cloze items that require articles seem to be of little value (unless an examiner wishes to test knowledge about articles specifically) compared to the other types of items.

Similarly, syntactic variation seems to have a great effect on both facility values and item-total correlations. There were a large number of answers that were semantically acceptable, but syntactically unacceptable. When those answers were accepted as correct, facility values for such items increased, quite dramatically in some cases (by even more than 100%). More importantly, item-total correlations also increased. This result can be explained by the fact that the concept of number is different in Japanese, and plural forms or verb endings relating to numbers do not exist. Therefore, it seems fairer—especially to Japanese learners—to accept syntactic variation as correct if the prime emphasis is on comprehension rather than on syntactic knowledge. Further study is needed to assess the significance of syntactic variation for native speakers of languages other than Japanese.

The preceding discussion of syntactic variation leads to the second point: When emphasis is on meaning rather than linguistic accuracy, cloze items seem to become more valid as a measure of reading comprehension. One of the problems with cloze tests is that they require not only reading ability but also productive skill. If it is reading comprehension that cloze tests are to measure, examiners may well be justified in accepting a different policy concerning answers that are syntactically incorrect, but nevertheless show that the reader has understood the meaning. The results with items that allow for alternative answers also support this point. Item-total correlations were low for items allowing alternative answers when the exact-word scoring method was applied, but they were as high as those for items that did not allow alternative answers when the acceptable-word scoring method was applied. This finding suggests that the results of a cloze test will better reflect the learners’ language ability when we accept answers that make sense in a given context.

Finally, I would like to mention the difficulty involved in categorizing items according to the kind of knowledge and the amount of context required to restore the word. This difficulty has been mentioned by the researchers who examined the cloze item characteristics (e.g., Abraham & Chapelle, 1992; Jonz, 1990). In this study, I have tried to go deeper into the issue by calling on the judgment of a number of experts. This process highlighted areas of difficulty in categorization.

In particular, content words were not straightforward. This difficulty may have arisen from the very nature of content words. Content words convey meanings, and each individual meaning contributes to a complicated semantic network in the text. At the same time, the meaning also acts on the readers’ schemata. It may be possible to restore deleted words by reading only one clause or sentence containing the blank, but other factors, such as co-textual knowledge (i.e., knowledge acquired by reading the preceding or following text or both) and general knowledge, also work together in the case of content words. It seems that answering cloze items, especially in the case of content words, requires more than one type of knowledge or amount of context. In the process of reading, this complicated interaction may occur automatically, but it is difficult to identify exactly what knowledge or how much context is required to fill in a blank correctly. To that extent, content word items in a cloze test appear to require more integrative ability than do function words (Oller, 1975). To facilitate further research in this area, it may be necessary to refine the frameworks for classification, as well as to ask more people to act as judges.

CONCLUSION
Summary and Implications

Detailed analysis of item characteristics in this study has demonstrated that cloze items vary in their difficulty and discrimination, as a function of the types of words deleted. There also appears to be a complex relationship between cloze item characteristics and scoring methods.

The preliminary investigation reported here showed that different starting points for deletion and different deletion rates might contribute to significant variations in the nature of cloze items produced from any one text. This finding raises questions concerning the consistency of fixed-ratio deletion cloze tests and contrasts with the
views of Bachman (1985) and Jonz (1990), who argue that different cloze versions are broadly consistent with each other. It would be wise to interpret the results of different cloze tests with caution, especially given that the variability of cloze item statistics seems to depend on the types of words deleted.

The overall findings of the research suggest that extra care is necessary in interpreting the results of cloze tests and research studies based on them, because the results—and the meaning of the scores—may have been affected by the particular set of words deleted and the scoring method used. We have at least to be aware of what kinds of words are involved in a particular cloze test so that we can understand and predict what the test results can tell us.

Limitations of the Study and Directions for Future Research

The present study had a number of limitations. They include: the number of items, the framework for item categorization, and the restricted range of the participants’ English proficiency. This prompts the following suggestions for future research:

1. Replication studies are needed with a greater number of cloze items, involving longer texts or more texts. Studies that involve relevant statistical analyses will also be useful in further determining the effects of scoring methods on the test performance of learners of different proficiency levels.

2. It would be desirable to elaborate or modify frameworks for identifying types of knowledge and the amount of context required to restore deleted words, especially to capture the nature of content-word items and to establish the validity of the frameworks with a greater number of judges.

3. Further studies concerning learners of different first languages and of a wider range of English-language proficiency may reveal different effects on cloze performance and the nature of cloze items.

Even though the present study was only exploratory in nature, its findings contribute to an understanding of the nature of cloze items and may enable researchers and test administrators to decide with more confidence which words should be deleted to achieve optimal language assessment and test construction. Above all, the analysis suggests ways in which the cloze test format, long regarded as not susceptible to analysis, can be studied by means of a finer-grained approach to reveal the different factors that underlie language proficiency.

ACKNOWLEDGMENTS

This paper developed ideas from an earlier article in Thames Valley University Working Papers in English Language Teaching (2, pp. 33–71). I would like to thank my colleagues and their students for their help in all stages of data collection. My thanks also go to Professor Peter Skehan and the anonymous reviewers for their helpful comments on the earlier drafts of this paper. Yuko Takakuwa provided invaluable assistance in data clearing during the original research. John Bray has been a constant source of expert advice and support.

NOTES

1 A T-unit is a way of identifying a unit of utterance. It is defined as “a main clause plus all subordinate clauses and nonclausal structures attached to or embedded in it” (Long & Sato, 1983, p. 284).

2 Fixed-ratio cloze is a cloze test in the original format (i.e., words are deleted at regular intervals) whereas rational-cloze is a modified format in which the test designer decides which words to delete.

3 The set of cloze tests was designed as part of a larger study that tested a research hypothesis to examine the effects of text organization and response format on reading comprehension performance (Kobayashi, 1995). The analysis presented here was conducted in the process.

4 See Kobayashi (1995) for details of text selection procedure. Twenty-seven “experts” (i.e., experienced EFL teachers) were invited to make judgments on the representativeness of text types, and two sets of four texts were finally selected on the basis of the degree of agreement among the experts.

5 A review of the previous studies suggests that the seventh word from the beginning or the seventh word after the first sentence or first two sentences seem to be common practice (e.g., Bachman, 1985; Chavez-Oller, Chihara, Weaver, & Oller, 1985; Jonz, 1991).

6 This evidence supported previous research findings (e.g., Alderson, 1979a, 1979b; Porter, 1978) and posed a serious question regarding the comparability of different cloze test versions. For example, 25 items from the same text might contain 6 content words and 19 function words from one starting point and 19 content words and 6 function words from another. Experiments with the texts also revealed that, among the content-word items, the number of words that appeared only once in the text could vary from 4 to 9 and the number of words that appeared more than once could vary from 2 to 10. This variation could cause problems. If the word category (i.e., function or content word) and the
number of occurrences in the text are predictors of item difficulties and reliabilities, it is then apparent that the variation in items will affect the quality of cloze tests. On the basis of this analysis, the starting point for deletion was decided for each text.

7 One was a native-speaker of Japanese who had been a teacher of English in Japan and had just completed her MA in applied linguistics. The other five were native speakers of English and had been involved in English language teaching, two of them having an MA in applied linguistics.

8 Six educated native speakers of English engaged in ELT were invited to answer the eight cloze tests. They were also asked to provide any alternative words that could fit in the blanks. These words were collected as “acceptable alternative” answers. In addition, one of the native speakers was continuously consulted when unlisted words appeared in the scripts with regard to the acceptability of the words.

REFERENCES


Sato, C. J. (1985). Task variation in interlanguage pho-
APPENDIX A
Sample Cloze Test

The industrialised countries between them possess 78% of all existing wealth. This means that the (1) countries, which are usually called the ‘Third World’, have about 22% of (2) wealth, even though their population is about 76% of the world’s total. (3) rich industrialised countries give aid to poorer Third World countries. However, this (4) sometimes does more harm than good. This is because many Western aid (5) are importing Western technology into the poorer countries. This has brought (6) two problems.

The first is that these Third World countries become dependent on (7) richer countries—they need them more and more. For example, a Third (8) country may be given expensive tractors. When the tractors go wrong, they (9) require skilled mechanics or expensive spare parts. Either way, the poor country (10) to pay money to the richer country.

Secondly, this kind of help (11) largely been based in the cities, making life there look more attractive. (12) people leave the countryside and move to the cities. As a result (13) have become overcrowded and there are all sorts of problems, from housing (14) to poor health facilities. At the same time the countryside becomes empty, (15) the country can no longer produce enough food for its people.

One (16) to these problems might be to encourage a different kind of technology. (17) technology should be based on what local people really need, and not (18) what people from outside think they need. It should be easy to (19), and it should help people in the countryside as well as the (20). Some countries are already using this kind of technology. Western aid organisations (21) teaching local people how to build simple machines using cheap, easily available (22) which can easily be replaced. They use these machines on their farms, (23) no longer need to go to the cities to get rich.

In (24) way, they learn to become independent of the West and Western technology. (25) the Chinese say: ‘Give a man a fish and feed him for a day. Teach him how to fish and he’ll feed himself for a lifetime.’

Key
1. other 2. the 3. Many 4. aid 5. organisations
6. about 7. the 8. World 9. will 10. needs
11. has 12. So 13. cities 14. shortages 15. and
16. solution 17. This 18. on 19. use 20. towns
21. are 22. materials 23. and 24 this 25. As

APPENDIX B
Samples of Acceptable Answers

Acc1 (both semantically and syntactically acceptable)
10. ought, has 12. Therefore 20. cities

Acc2 (semantically acceptable, but syntactically unacceptable)
11. have 14. shortage 20. city, town
**APPENDIX C**

Descriptive Statistics

The table below shows the means (M), standard deviations (SD), reliability (α), and correlations with the proficiency test (r) of the eight texts by the three different scoring methods.

<table>
<thead>
<tr>
<th></th>
<th>Ex</th>
<th>Acc1</th>
<th>Acc2</th>
<th>Ex</th>
<th>Acc1</th>
<th>Acc2</th>
<th>Ex</th>
<th>Acc1</th>
<th>Acc2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text 1 (n = 25)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>7.94</td>
<td>11.5</td>
<td>11.9</td>
<td>7.89</td>
<td>9.68</td>
<td>10.2</td>
<td>15.8</td>
<td>21.2</td>
<td>22.1</td>
</tr>
<tr>
<td>SD</td>
<td>2.77</td>
<td>3.75</td>
<td>3.81</td>
<td>3.66</td>
<td>4.43</td>
<td>4.41</td>
<td>5.64</td>
<td>7.67</td>
<td>7.66</td>
</tr>
<tr>
<td>α</td>
<td>0.62</td>
<td>0.72</td>
<td>0.72</td>
<td>0.75</td>
<td>0.80</td>
<td>0.79</td>
<td>0.80</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>r</td>
<td>.59**</td>
<td>.66**</td>
<td>.67**</td>
<td>.54**</td>
<td>.57**</td>
<td>.54**</td>
<td>.64**</td>
<td>.66**</td>
<td>.64*</td>
</tr>
<tr>
<td><strong>Text 2 (n = 25)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>6.91</td>
<td>10.7</td>
<td>10.9</td>
<td>7.35</td>
<td>9.15</td>
<td>9.53</td>
<td>14.3</td>
<td>19.8</td>
<td>20.5</td>
</tr>
<tr>
<td>SD</td>
<td>3.32</td>
<td>4.65</td>
<td>4.76</td>
<td>3.39</td>
<td>4.23</td>
<td>4.21</td>
<td>6.42</td>
<td>8.56</td>
<td>8.60</td>
</tr>
<tr>
<td>α</td>
<td>0.72</td>
<td>0.80</td>
<td>0.81</td>
<td>0.73</td>
<td>0.80</td>
<td>0.79</td>
<td>0.85</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>r</td>
<td>.70**</td>
<td>.75**</td>
<td>.76**</td>
<td>.73**</td>
<td>.73**</td>
<td>.72**</td>
<td>.75**</td>
<td>.78**</td>
<td>.77*</td>
</tr>
<tr>
<td><strong>Text 3 (n = 25)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>8.31</td>
<td>9.36</td>
<td>9.98</td>
<td>7.77</td>
<td>9.93</td>
<td>10.3</td>
<td>16.3</td>
<td>19.3</td>
<td>20.4</td>
</tr>
<tr>
<td>SD</td>
<td>4.56</td>
<td>4.92</td>
<td>4.97</td>
<td>3.58</td>
<td>4.40</td>
<td>4.60</td>
<td>7.74</td>
<td>8.97</td>
<td>9.11</td>
</tr>
<tr>
<td>α</td>
<td>0.82</td>
<td>0.84</td>
<td>0.83</td>
<td>0.73</td>
<td>0.80</td>
<td>0.80</td>
<td>0.88</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>r</td>
<td>.73**</td>
<td>.75**</td>
<td>.74**</td>
<td>.73**</td>
<td>.75**</td>
<td>.77**</td>
<td>.77**</td>
<td>.79**</td>
<td>.79*</td>
</tr>
<tr>
<td><strong>Text 4 (n = 25)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>6.52</td>
<td>9.94</td>
<td>10.3</td>
<td>5.29</td>
<td>7.33</td>
<td>7.89</td>
<td>10.8</td>
<td>15.9</td>
<td>17.6</td>
</tr>
<tr>
<td>SD</td>
<td>3.36</td>
<td>4.30</td>
<td>4.78</td>
<td>2.65</td>
<td>3.73</td>
<td>3.84</td>
<td>5.57</td>
<td>7.66</td>
<td>8.13</td>
</tr>
<tr>
<td>α</td>
<td>0.73</td>
<td>0.79</td>
<td>0.83</td>
<td>0.62</td>
<td>0.77</td>
<td>0.78</td>
<td>0.82</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>r</td>
<td>.64**</td>
<td>.74**</td>
<td>.73**</td>
<td>.64**</td>
<td>.67**</td>
<td>.64**</td>
<td>.69**</td>
<td>.75**</td>
<td>.73*</td>
</tr>
</tbody>
</table>

Note. N = Number of subjects; n = Number of items. Ex = Exact-word scoring method; Acc1 = Semantically and syntactically acceptable-word scoring method; Acc2 = Semantically acceptable, but syntactically unacceptable-word scoring method.

*p < .01. **p < .001.

---

**APPENDIX D**

Reliability of Content and Function Word Items

<table>
<thead>
<tr>
<th></th>
<th>Content Words</th>
<th>Function Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 1</td>
<td>.44</td>
<td>.61</td>
</tr>
<tr>
<td>Text 2</td>
<td>.54</td>
<td>.66</td>
</tr>
<tr>
<td>Text 3</td>
<td>.72</td>
<td>.75</td>
</tr>
<tr>
<td>Text 4</td>
<td>.56</td>
<td>.65</td>
</tr>
<tr>
<td>Text 5</td>
<td>.64</td>
<td>.72</td>
</tr>
<tr>
<td>Text 6</td>
<td>.29</td>
<td>.45</td>
</tr>
<tr>
<td>Text 7</td>
<td>.47</td>
<td>.67</td>
</tr>
<tr>
<td>Text 8</td>
<td>.43</td>
<td>.65</td>
</tr>
</tbody>
</table>
APPENDIX E
Correlation Coefficients between the Cloze Test Results Scored by Three Different Scoring Methods

<table>
<thead>
<tr>
<th>Text 1</th>
<th>Text 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex</td>
<td>Acc1</td>
</tr>
<tr>
<td>Ex</td>
<td>1.00</td>
</tr>
<tr>
<td>Acc1</td>
<td>.914</td>
</tr>
<tr>
<td>Acc2</td>
<td>.900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text 2</th>
<th>Text 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex</td>
<td>Acc1</td>
</tr>
<tr>
<td>Ex</td>
<td>1.00</td>
</tr>
<tr>
<td>Acc1</td>
<td>.984</td>
</tr>
<tr>
<td>Acc2</td>
<td>.966</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text 3</th>
<th>Text 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex</td>
<td>Acc1</td>
</tr>
<tr>
<td>Ex</td>
<td>1.00</td>
</tr>
<tr>
<td>Acc1</td>
<td>.930</td>
</tr>
<tr>
<td>Acc2</td>
<td>.918</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text 4</th>
<th>Text 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex</td>
<td>Acc1</td>
</tr>
<tr>
<td>Ex</td>
<td>1.00</td>
</tr>
<tr>
<td>Acc1</td>
<td>.917</td>
</tr>
<tr>
<td>Acc2</td>
<td>.879</td>
</tr>
</tbody>
</table>

*Note.*
1. Ex = exact-word scoring method; Acc1 = semantically and syntactically acceptable-word scoring method; Acc2 = semantically only acceptable-word scoring method.
2. All the values presented here are significant at $p < .001$.

Forthcoming in *The Modern Language Journal*


Christiane Bongartz & Melanie Schneider. “Linguistic Development in Social Contexts: A Study of Two Brothers Learning German”

Bryan Smith. “Computer-Mediated Negotiated Interaction: An Expanded Model” (Held over from MLJ 86,4 for page limitations)

Deborah Arteaga, Julia Herschensohn, & Randall Gess. “Focusing on Phonology to Teach Morphological Form in French”

Julie Belz & Andreas Muller-Hartmann. “Teachers Negotiating German-American Telecollaboration along the Institutional Fault Line”