TAILORING CLOZE: THREE WAYS TO IMPROVE CLOZE TESTS

JAMES DEAN BROWN
University of Hawai‘i
AMY D. YAMASHIRO
Nihon University
ETHEL OGANE
Meiji University

Cloze procedure first surfaced when Taylor (1953) investigated its effectiveness as a tool for measuring the readability of written materials. Research soon turned to cloze as a measure of native-speaker reading proficiency (Ruddell, 1964; Gallant, 1965; Bormuth, 1965, 1967; Crawford, 1970). In the 1960s, studies began to surface on cloze as a measure of overall ESL proficiency (for overviews of this L2 research on cloze, see Alderson, 1978 or Oller, 1979). However, a careful review of the L2 cloze literature in the last three decades reveals results that are far from consistent in terms of their reliability and validity.

Studies of cloze test reliability address the issue of the consistency of cloze scores. Reliability statistics can range from a minimum of .00 (for completely inconsistent scores) to a maximum of 1.00 (for completely consistent scores). To date, the research has shown cloze test reliabilities that have ranged from .31 to .96 (Darnell, 1970; Oller,

We must stress that, throughout this paper, we are studying ways of improving cloze tests for norm-referenced purposes (e.g., admissions or placement testing) rather than for criterion-referenced purposes (e.g., diagnostic, progress, or achievement testing). These two sets of purposes are quite different as explained in Brown (1988b, 1995, 1996), and improving cloze tests for criterion-reference purposes would require quite different strategies.

1972a; Pike, 1973; Jonz, 1976; Alderson, 1979; Mullen, 1979; Hinofotis, 1980; Bachman, 1985; Brown, 1980, 1983, 1984, 1988a, 1989, 1994), which means that the cloze tests have ranged from 31% reliable to 96%.

Studies of cloze validity address the degree to which a cloze test is measuring what it claims to be measuring, which in most studies means overall English language proficiency. One common approach to investigating cloze validity is called the criterion-related validity strategy, which involves calculating a correlation coefficient between the results on a cloze test and parallel results on some other, well-respected criterion measure of overall English language proficiency such as the TOEFL. The coefficient of determination, which is obtained by squaring the correlation coefficient, shows the percentage of overlapping variance between the cloze procedure scores and whatever criterion measure scores are involved. The criterion-related validity coefficients reported in various studies have ranged from .43 to .91 (Conrad, 1970; Darnell, 1970; Oller & Inal, 1971; Oller, 1972a & b; Irvine, Atai, & Oller, 1974; Stubbs & Tucker, 1974; Mullen, 1979; Alderson, 1979, 1980; Hinofotis, 1980; Bachman, 1985; Brown, 1980, 1984, 1988a; Revard, 1990). Since the corresponding coefficients of determination are .19 to .83, these studies indicate that the relationships of various cloze tests to criterion measures have ranged from very low relationships (19%) to fairly strong relationships (83%).

In fact, a single recent study not included in the above discussion (Brown, 1993, in which 50 randomly selected passages from a public library were made them into cloze tests and randomly administered to 2298 students at 18 universities in Japan) found that the 50 cloze tests varied considerably in reliability (.17 to 1.00 for different passages and different ways of estimating reliability) and validity (with criterion-related validity coefficients ranging from .00 to .71). Indeed, overall, the 50 cloze tests were not particularly reliable or valid.

In one way or another, the research cited above investigated the best ways to develop reliable and valid cloze tests in terms of at least the following variables: (a) scoring scheme, (b) deletion patterns (e.g., every 5th word, every 7th word, etc.), (c) number of items, (d) blank lengths, (e) passage difficulties, and (f) native versus non-native performance.

Brown (1984) argues that the single most important variable in creating reliable and valid cloze tests may be how well a given passage fits the particular group of students being tested. How well a cloze test fits a particular group of students may be a function
of any or all of the six variables discussed in the previous paragraph. In addition, Brown argues that the effectiveness of any blending of those variables is strongly related to the degree to which the cloze test produces centered scores for the people being tested (as indicated by the mean) and widely dispersed scores (as shown by the standard deviation). Brown further suggests that three methods exist for creating a well-centered cloze test with widely-dispersed scores:

1. **The hit-or-miss method** - This method requires administering a number of cloze tests in order to find the one that is functioning best.

2. **The modification method** - This method requires administering a cloze test, analyzing the results, and modifying variables like scoring scheme, deletion pattern, or number of items to improve how well it functions.

3. **The tailored-cloze method** - This method employs traditional item analysis techniques to improve how well a cloze test functions. Brown (1988a) showed how the tailored-cloze method can be applied.

The purpose of the present study was to explore how the hit-and-miss, modification, and tailored-cloze methods can be used to increase the efficiency and effectiveness of cloze tests. To help organize the results of the present study, the following research questions were posed:

1. What is the effect on the mean and standard deviation for a cloze test when applying the hit-or-miss, modification, and tailored-cloze methods?

2. To what degree are the item facilities and discrimination indices changed by applying the hit-or-miss, modification, and tailored-cloze methods?

3. What is the effect on reliability of applying the hit-or-miss, modification, and tailored-cloze methods?

A much fuller description of the steps involved in using the hit-or-miss, modification, and tailored-cloze methods will be provided later in this paper.

**METHODS**

**Participants**

The participants in this study were all students in a highly-ranked private secondary school in Japan that provides automatic entrance to a top-tier university in Japan. About 30% of the matriculated students are *returnees*, students who have lived and studied overseas for at least two years because their parents were working there. The returnees
are pulled out and taught in separate English classes. Regular students are tracked into regular and higher-proficiency levels. Regular students may elect or be recommended to join the returnee class for extra challenge in their English studies.

Three first-year returnee classes of about 25 students, and three second-year returnee classes of about 25 students each were used in this study. About 10% were high-proficiency regular Japanese students who either volunteered or were recommended to join the returnee classes. At the end of the school year, all first-year students take the Pre-TOEFL and roughly two-thirds of the returnee class receive perfect scores of 500, while the remaining students score within the mid-to-high 400s. At the same time, all second-year students take the institutional TOEFL test, and the returnee class students typically average about 570, with scores ranging from the high 400s to the high 600s.

Materials

The passage used for this study, entitled “The Science of Automatic Control,” first appeared in Bachman (1985). Five cloze tests, including Bachman’s original fixed-ratio cloze, were used in this study as follows: Form A in this study was identical to Bachman’s original fixed-ratio cloze test; Form B deleted the word one to the right of the original deletion; Form C deleted the word two to the right of the original deletion; Form D deleted the word one to the left; and Form E deleted the word two to the left of the original deletion (see Appendix A for example directions and the first ten items of this form; see Appendix B for the answer key for those first ten items).

The cloze tests were scored in two ways: exact-answer scoring and acceptable-answer scoring. The exact-answer scoring only counts as correct that word which was originally deleted from the blank. The acceptable-answer scoring involves creating a glossary of possible answers for each blank. The glossary was created by the three researchers, all native speakers of English, who generated an answer key (with a glossary for each item) for each of the five forms before the tests were scored. Additional acceptable answers were added, after reaching agreement between two of the researchers during the scoring phase; all previously marked tests were then rechecked for the newly added acceptable answers.
TAILORING CLOZE: THREE WAYS TO IMPROVE CLOZE TESTS

Procedures

The cloze tests were administered in two waves. The first wave was administered in September 1997 (near the start of the second term); while the second wave was administered in February 1998 (in the middle of the third term).

First wave administration. The five different versions of the cloze test were piled in repeating patterns of Forms A, B, C, D, and E. The test administrators, two teachers from the school, were instructed to pass out the tests proceeding from the top of the pile to the bottom for each class. The teachers allowed the students 15 minutes of class time to take the cloze tests. The participants were told that the scores would not affect their course grades.

First wave item analysis. Each item on each cloze test was then separately scored for exact and acceptable answers. The data were input into a separate computer spreadsheet for each form and scoring scheme—using 1s to represent correct answers and 0s for incorrect ones. Then total scores were calculated for each student and the data were sorted in descending order by total scores. Overall item facility values were calculated for each item. Item facility values were also computed separately for the top third of students (that is, the highest third in terms of their total scores) and bottom third (that is, the lowest third in terms of total scores) for each item. Finally, item discrimination indexes were calculated by subtracting the item facility for the top third minus the item facility for the bottom third for each item.

The item statistics for the exact-answer scoring of the five original cloze tests were examined, and the 30 items with the highest discrimination values were used to create one new form labeled Revised EX here. Then the item statistics for the acceptable-answer scoring of the five original cloze tests were examined, and the 30 items with the highest discrimination values were used to create a second new form labeled Revised AC in this study. In cases where equal discrimination values were found, the item that had not already been used or was less frequently represented elsewhere in the blanks was selected.

Second wave administration. About five months later, the Revised EX and Revised AC were placed in a pile in alternating manner, and again, each teacher was instructed to pass out the tests from the top continuing straight through the pile for each class. The number of participants differed from the first wave to the second due to absences. As in the first wave administration, each teacher allowed 15 minutes for administering the
cloze tests, and the participants were told that the scores would not affect their course grades.

Second wave item analysis. The same procedures described above were used for entering the data and for calculating the item facility and discrimination values in spreadsheet programs. Revised EX was scored for exact answers and Revised AC was scored for acceptable answers. A few additional acceptable answers were added during the scoring phase for this wave with all previously marked tests in the second wave being rechecked for the newly added acceptable answers.

RESULTS

Recall that the purpose of this study was to explore how the hit-and-miss, modification, and tailored-cloze methods can be used to increase the efficiency and effectiveness of cloze tests. Like the tests, the results of this study will be reported in a first wave (the five original cloze versions) and second wave (the revised versions developed from what we learned in the pilot testing).

First Wave: Five Original Cloze Versions

In this subsection, we will cover three types of statistics as they apply to the five original cloze versions in this study: descriptive statistics, item analyses, and reliability estimates. They will be defined and discussed in that order.

Descriptive statistics. In this study, the purpose of the descriptive statistics is to present a picture of the distribution of scores on each test. The descriptive statistics included here are the mean, standard deviation, number of students, low score, high score, and range.

For the purposes of this study, the mean ($M$) will be defined simply as the arithmetic average. Notice, in Table 1, that the means for each form are generally several points higher for the acceptable-answer scoring than for the exact-answer scoring. This pattern is to be expected given that the exact answers are included in the acceptable-answer key along with other possibilities. As a result of the acceptable-answer cloze versions having higher means, they are also better centered, that is, they are closer to the middle score of 15, half-way along the scale of possible scores from 0 to 30. An ideal norm-referenced test will be well-centered and thus will stand a good chance of efficiently spreading the
students out into a normal distribution. Hence the acceptable-answer scoring, which appears to be better centered, may be more efficient for norm-referenced purposes.

That view is re-enforced by the fact that the standard deviations (SD) reported in Table 1 for the original acceptable-answer cloze tests are systematically higher than those reported for the corresponding forms scored for exact answers. The standard deviation "provides a sort of average of the differences of all scores from the mean" (Brown, 1988a, p. 69), which indicates the degree to which a test is efficiently spreading the students out. For example, the standard deviation for Form AC A (scored for acceptable answers) is 4.12, which means that on average the students were 4.12 points above or below the mean; of course, some students were more than 4.12 points above or below the mean, but on average it was 4.12 points. In contrast, the standard deviation for Form EX A (scored for exact answers) is only 3.41. The standard deviations for the acceptable-answer scoring are consistently higher than those for the exact-answer scoring for each of the five forms. Thus the acceptable-answer scoring appears to systematically be spreading the students out more efficiently than the exact-answer scoring.

<table>
<thead>
<tr>
<th>SCORING FORM</th>
<th>M</th>
<th>SD</th>
<th>n</th>
<th>LOW</th>
<th>HIGH</th>
<th>RANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXACT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX A</td>
<td>8.59</td>
<td>3.41</td>
<td>29</td>
<td>1</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>EX B</td>
<td>9.33</td>
<td>3.76</td>
<td>30</td>
<td>3</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>EX C</td>
<td>9.59</td>
<td>3.89</td>
<td>29</td>
<td>4</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>EX D</td>
<td>7.79</td>
<td>3.11</td>
<td>28</td>
<td>2</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>EX E</td>
<td>11.44</td>
<td>3.39</td>
<td>27</td>
<td>6</td>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>ACCEPTABLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC A</td>
<td>10.90</td>
<td>4.12</td>
<td>29</td>
<td>2</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>AC B</td>
<td>12.23</td>
<td>3.86</td>
<td>30</td>
<td>5</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>AC C</td>
<td>14.76</td>
<td>5.19</td>
<td>29</td>
<td>6</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>AC D</td>
<td>11.32</td>
<td>4.35</td>
<td>28</td>
<td>4</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>AC E</td>
<td>15.26</td>
<td>4.22</td>
<td>27</td>
<td>8</td>
<td>27</td>
<td>20</td>
</tr>
</tbody>
</table>
The number of students (n) on each of the forms ranges from 27 for Forms EX E and AC E to 30 for Forms EX B and AC B. Since the students were randomly assigned to forms, these fluctuations were probably due solely to chance factors. Generally, 28 to 30 students is considered the minimum acceptable sample for small-scale pilot testing.

Finally, the low score and high score are just what they suggest, the lowest and highest scores attained by the students. The range is calculated by subtracting the low score from the high score and adding 1 (range = high - low + 1). [Note that the 1 is added so the scores at both ends of the range are included.] For example, the low, high, and range for the Form EX A are 1, 16, and 16, respectively, which means that the scores extend from 1 to 16 with a range of 16 points (range = high - low + 1 = 16 - 1 + 1 = 16). Notice also that, like the standard deviations, the ranges are higher in magnitude for the acceptable-answer scoring on all five forms than for the exact-answer scoring. This, combined with the standard deviation results already discussed, provides further evidence that the acceptable-answer scoring is systematically spreading students out more efficiently than the exact-answer scoring.

Descriptive statistics can be used to apply the hit-or-miss method of cloze development. Recall that the hit-or-miss method requires administering a number of cloze tests and selecting the one that works best. Examining just the five cloze tests scored for exact answers, the descriptive statistics in Table 1 indicate that Form EX E has the best centered mean, while EX C has the largest standard deviation, and EX B has the highest range. Thus based solely on the exact-answer scoring, we cannot definitively decide on a single best cloze using the hit-or-miss method, though three of them would be candidates from one point of view or another.

Descriptive statistics can also be used to apply the modification method. Remember that the modification method requires administering a cloze test, analyzing the results, and modifying the variables like scoring method, deletion pattern, or number of items to improve how well it functions. One way to increase the effectiveness of a norm-referenced test is to improve how the test is centering and dispersing scores. In this case, the focus was on modifying the scoring method from the exact-answer to the acceptable-answer in order to improve the cloze tests in terms of means, standard deviations, and ranges. The results indicate that, in all cases, the acceptable-answer scoring created scores that are better centered (as indicated by the fact that the means are closer to the midpoint of 15 on the 0 to 30 scale of possible scores) and better dispersed (as indicated by the generally higher magnitude of the standard deviations and ranges).
TAILORING CLOZE: THREE WAYS TO IMPROVE CLOZE TESTS

Combining both the hit-or-miss method with the modification method, Forms AC C and AC E appear to have about equally well-centered means, while AC C has the highest standard deviation and AC E has the highest range. Thus even though there are clearly two front runners, based solely on the descriptive statistics, we still cannot decide definitively which single form is best overall in these five original cloze versions scored for acceptable-answers.

Item analyses. Item analyses are a set of statistical techniques used to assess the effectiveness of individual items on a test. In this study, we will consider three norm-referenced item statistics: item facility, item discrimination, and the number of zero items.

Item facility (IF) is simply the proportion of students who answered a particular item correctly. Thus, if no student answered an item correctly, the item facility would be .00 while, if all students answered correctly, it would be 1.00. The range of possible IF values is from .00 to 1.00. In a case where 14 out of 24 students answered correctly, the IF would be 14/24 = .58333, or about .58. By moving the decimal point two places to the right, the IF value can be interpreted as a percentage. Thus for the .58 example value, it would be correct to say that 58% of the students answered correctly. For norm-referenced purposes items with IF values between .30 and .70 with a mean of .50 are often considered ideal. The IF values reported in Table 2 are mean IFs calculated across the 30 items on each form. Notice that the mean IF values for the acceptable-answer scoring are in general closer to the ideal value of .50 than the ones for the exact-answer scoring.

Item discrimination (ID) was calculated in this case by subtracting the item facility for the upper third of students (ordered from high to low according to their total test scores) minus the item facility for the lower third of students. For example, if 80% of the top one-third of students answered correctly, and only 20% of the bottom one-third answered correctly, the ID would be ID = IF_{top} - IF_{bottom} = .80 - .20 = .60. ID can range from .00 to 1.00 in a positive direction (if high students tend to answer correctly more often than low students) and from .00 to -1.00 in a negative direction (if low students tend to answer correctly more often than high students). Thus, generally, the higher the ID value the better because high values indicate items that discriminate well for norm-referenced purposes. Notice in Table 2 that the mean ID values are equal to or higher for the acceptable-answer scoring than for the exact-answer scoring.
Table 2

**Item Statistics for Five Original Cloze Versions and Two Scoring Methods**

<table>
<thead>
<tr>
<th>SCORING FORMS</th>
<th>MEAN IF</th>
<th>MEAN ID</th>
<th># OF 0 ITEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EXACT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX A</td>
<td>0.29</td>
<td>0.23</td>
<td>4</td>
</tr>
<tr>
<td>EX B</td>
<td>0.31</td>
<td>0.26</td>
<td>9</td>
</tr>
<tr>
<td>EX C</td>
<td>0.32</td>
<td>0.29</td>
<td>6</td>
</tr>
<tr>
<td>EX D</td>
<td>0.26</td>
<td>0.23</td>
<td>8</td>
</tr>
<tr>
<td>EX E</td>
<td>0.38</td>
<td>0.22</td>
<td>4</td>
</tr>
<tr>
<td><strong>ACCEPTABLE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC A</td>
<td>0.36</td>
<td>0.28</td>
<td>2</td>
</tr>
<tr>
<td>AC B</td>
<td>0.41</td>
<td>0.26</td>
<td>4</td>
</tr>
<tr>
<td>AC C</td>
<td>0.49</td>
<td>0.38</td>
<td>1</td>
</tr>
<tr>
<td>AC D</td>
<td>0.38</td>
<td>0.30</td>
<td>2</td>
</tr>
<tr>
<td>AC E</td>
<td>0.51</td>
<td>0.29</td>
<td>0</td>
</tr>
</tbody>
</table>

The *number of zero items* (# OF 0 ITEMS) is a simple tally of the number of items that nobody answered correctly. These are items that are contributing absolutely nothing to the efficiency of a norm-referenced test because zero percent of students answered correctly and ID would also be zero. Notice in Table 2 that the number of such items is considerably smaller for the acceptable-answer forms than for the corresponding exact-answer forms.

Combining the hit-or-miss and modification methods, we note that Form AC C has the highest overall ID value at .38, an IF value very close to the ideal of .50 at .49, and only one non-functioning item that nobody answered correctly. Thus, based on the descriptive statistics and item analyses so far, Form AC C appears to be the best overall cloze test in these pilot data.
Table 3
Reliability Statistics for Five Original Cloze Versions and Two Scoring Methods

<table>
<thead>
<tr>
<th>SCORING FORMS</th>
<th>k</th>
<th>ALPHA</th>
<th>SPLIT-HALF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXACT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX A</td>
<td>30</td>
<td>0.645</td>
<td>0.636</td>
</tr>
<tr>
<td>EX B</td>
<td>30</td>
<td>0.757</td>
<td>0.803</td>
</tr>
<tr>
<td>EX C</td>
<td>30</td>
<td>0.751</td>
<td>0.784</td>
</tr>
<tr>
<td>EX D</td>
<td>30</td>
<td>0.640</td>
<td>0.568</td>
</tr>
<tr>
<td>EX E</td>
<td>30</td>
<td>0.636</td>
<td>0.445</td>
</tr>
<tr>
<td>ACCEPTABLE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC A</td>
<td>30</td>
<td>0.738</td>
<td>0.779</td>
</tr>
<tr>
<td>AC B</td>
<td>30</td>
<td>0.715</td>
<td>0.772</td>
</tr>
<tr>
<td>AC C</td>
<td>30</td>
<td>0.832</td>
<td>0.814</td>
</tr>
<tr>
<td>AC D</td>
<td>30</td>
<td>0.768</td>
<td>0.712</td>
</tr>
<tr>
<td>AC E</td>
<td>30</td>
<td>0.718</td>
<td>0.612</td>
</tr>
</tbody>
</table>

Reliability estimates. Table 3 provides information about the reliability of the five original versions of the cloze tests. As explained earlier, reliability estimates range from .00 to 1.00, that is, from no reliability to complete reliability. By moving the decimal point two places to the right, reliability estimates can be interpreted as percentages. Thus a reliability estimate of .86 would indicate that a test is 86% reliable and, by extension, 14% unreliable.

The first column of numbers in Table 3 gives the number of items (k). In all cases, there were 30 items. This information is important only insofar as it indicates that the length of each cloze test was the same. Since reliability and test length are related, we chose to hold this variable constant.

The second column of numbers shows the Cronbach alpha estimates of reliability, while the third column shows the split-half estimates (adjusted for full-test reliability using the Spearman-Brown formula) for each of the forms (for complete explanations of both Cronbach and split-half estimates, see Brown, 1996). Notice that, in most cases, the two reliability estimates are fairly similar. Examining both estimates at the same time for the exact-answer scoring, Form EX B has the highest reliability estimates. Combined
with the fact that it had the second highest standard deviation and the highest range, as well as the second highest mean ID, this reliability information might lead us to select EX B as the best of the five exact-answer original cloze tests if we were using the hit-or-miss method for the exact-answer forms only.

Turning to the modification method, notice that the reliability estimates in Table 3 also indicate that, with only one exception, the acceptable-answer scoring produced higher reliability estimates than the corresponding exact-answer form.

Combining both the hit-or-miss method with the modification method, Form AC C has the highest reliability estimates. That plus the fact that AC C has the second best centered mean, the highest standard deviation, and a reasonably high range, might justifiably lead us to select AC C as the best overall cloze test among the five original forms and two scoring schemes.

**Second Wave: Revised Cloze Versions**

Recall that the item analysis statistics for the exact- and acceptable-answer results on the five original cloze versions were used to select items and create two revised cloze versions, that is, Revised EX based on the original exact-answer results and Revised AC based on the original acceptable-answer results (as explained in the Materials and Procedures sections above). In this subsection, we will focus on the results of these two revised cloze tests in terms of the same three types of statistics discussed previously for the five original cloze versions: descriptive statistics, item analyses, and reliability estimates.

**Descriptive statistics.** In Table 1, we saw that the means for the acceptable-answer scoring of the five original cloze tests were several points higher than for the exact-answer scoring and were also somewhat better centered for the acceptable-answer scoring and by extension more efficient for norm-referenced purposes. The descriptive statistics for the revised cloze versions (see Table 4) indicate that the means for the revised cloze tests are higher than those for any of the original cloze versions in either scoring scheme. In addition, the standard deviations and ranges are considerably larger for the revised versions than for the original versions. Between the two revised versions, the descriptive statistics indicate that Revised AC is reasonably well-centered and better at dispersing the students than any of the five original versions and even somewhat better than Revised EX.
**Descriptive Statistics for Two Revised Cloze Versions (N = 144)**

<table>
<thead>
<tr>
<th>FORM</th>
<th>M (SD)</th>
<th>n</th>
<th>LOW</th>
<th>HIGH</th>
<th>RANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revised EX</td>
<td>15.31 (5.30)</td>
<td>71</td>
<td>5</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>Revised AC</td>
<td>16.51 (6.29)</td>
<td>73</td>
<td>3</td>
<td>27</td>
<td>25</td>
</tr>
</tbody>
</table>

**Item analyses.** The item analysis statistics show very similar results. The mean IFs for the acceptable-answer scoring of the five original cloze versions (shown in Table 2) were considerably higher than for the exact-answer scoring and overall were somewhat better centered. Acceptable-answer scoring once again appears to be more efficient for norm-referenced testing purposes. The item analyses for the revised cloze versions (shown in Table 5) indicate that the IFs for the revised cloze tests are higher than those for any of the original cloze versions in the same scoring scheme. In addition, the IDs are considerably larger for the revised versions than for the corresponding scoring schemes in the original versions. Indeed, the IDs for the revised versions can be said to be comparable or higher than the IDs for all the original versions. Notice also in Table 5 that none of the items in the Revised EX or Revised AC were zero items, while all but one of the forms shown in Table 2 had at least one and as many as nine such non-functioning items. Between the revised versions, the item analyses indicate that Revised AC is reasonably well-centered and better at discriminating among the students than any of the five original versions and somewhat better than Revised EX.

**Reliability estimates.** Table 6 shows the Cronbach alpha and split-half reliabilities for the revised cloze versions. In each case, the two estimates are approximately the same. Examining both at the same time, the reliabilities reported in Table 6 are higher in all
cases than any of the reliabilities for the corresponding scoring scheme in Table 3. Thus generally, tailoring the cloze appears to have improved the reliability of these tests. In addition, the acceptable-answer reliabilities in Table 6 are both somewhat better than those for the exact-answer scoring and better than any of the reliability estimates for the five original versions reported in Table 3.

Table 6
Reliability Statistics for Two Revised Cloze Versions

<table>
<thead>
<tr>
<th>FORM</th>
<th>k</th>
<th>ALPHA</th>
<th>SPLIT-HALF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revised EX</td>
<td>30</td>
<td>0.828</td>
<td>0.831</td>
</tr>
<tr>
<td>Revised AC</td>
<td>30</td>
<td>0.869</td>
<td>0.845</td>
</tr>
</tbody>
</table>

Combined with what we learned from the descriptive statistics and from the item analyses, we would have to say that, on the whole, the revised cloze version using acceptable-answer scoring would be the best cloze test to surface in this study.

**DISCUSSION**

In order to provide direct answers to the research questions posed at the beginning of this study, we will use them as headings for the three sections in this Discussion section.

1. **What Is the Effect on the Mean and Standard Deviation for a Cloze Test When Applying the Hit-or-Miss, Modification, and Tailored-Cloze Methods?**

   In this study, using the shot-gun approach that Brown (1984) called the hit-or-miss method involved administering five cloze tests and selecting the best of the five exact-answer forms for future administrations. In this particular investigation, that turned out to be more complicated than we had anticipated. Depending on the criterion used, three different cloze passages might be chosen:
   1. Based on the best centered mean (that is, the mean closest to the middle on the continuum from 0 to 30 points), Form EX E should be selected (see Table 1);
   2. Based on the highest standard deviation, Form EX C should be picked (see Table 1);
   3. Based on the largest range, Form EX B should be chosen (see Table 1).
We expected those three criteria to converge on a single passage, but they did not. If they had, the choice of which passage to keep for future administrations would have been considerably simplified. However, even as the results turned out, a teacher might decide on one passage or another based on one of the three criteria listed above, depending on what is most important in the particular situation. For instance, it might be more important in a particular situation (where students’ grades depended in part on the scores) to have a passage that is easier and well centered than to have a passage that is spreading students out. In that case, criterion a. above might be most important and thus EX E would be selected. In another situation, where important placement decisions would be made based on the cloze test, spreading the students out might be more important. In this second case, criterion b. above might be most important, and thus, EX C would appropriately be selected.

Elsewhere in this study, the modification method, as it was labeled by Brown (1984), required changing variables like scoring scheme, deletion pattern, or number of items to improve how well a cloze test functions. Modifying the scoring scheme clearly had a salutary effect on the descriptive statistics for the cloze tests in this study. As shown in Table 1, the means and standard deviations were higher in all cases for the acceptable-answer scoring than for the exact-answer scoring, and the ranges were equal or higher in all cases. Thus, at least this one type of modification seems to have improved the descriptive characteristics of all five versions of the cloze tests under investigation here.

Combining both the hit-or-miss method with the modification method, Forms AC C and AC E appear to be about equally well centered in terms of their means, and AC E has the highest overall range (as shown in Table 1), while AC C has the highest standard deviation. Thus based solely on the descriptive statistics for the hit-or-miss and modification methods, we could not decide definitively which single form is best overall in these five original cloze versions scored for acceptable-answers.

The tailored-cloze method, as defined in Brown (1984, 1988a) and as used here, required using traditional item analysis techniques to improve how well a cloze passage functions. In the present study, the two cloze versions that were tailored in that manner proved to be far superior to any of the original versions in terms of having higher standard deviations and ranges (as a quick comparison of Tables 1 and Table 4 will show).

However, based solely on the descriptive statistics, we cannot decide definitively, using either the tailored-cloze method alone, which of the cloze tests is likely to function
the best in future administrations. At best, we can say that Revised EX is the best centered, while Revised AC is very well centered and has the highest range and standard deviation.

2. To What Degree Are the Item Facilities and Discrimination Indices Changed by Applying the Hit-or-Miss, Modification, and Tailored-Cloze Methods?

Turning to the item analyses and the hit-or-miss method, the selection of which test works best from among the five exact-answer original cloze versions would differ depending on the criteria used in a manner very similar to the descriptive statistics and for similar reasons. If the mean IF is the primary concern, EX E should be selected. If the mean ID is the primary concern, EX C should be selected. If the number of non-functioning items is the primary concern, EX A or EX E should be selected.

As for applying the modification method, we can certainly say that, similar to the descriptive statistics, modifying the scoring scheme clearly had a salutary effect on the item statistics for the cloze tests in this study. As shown in Table 2, in all cases, the mean IFs and mean IDs for the acceptable-answer scoring were equal to or higher than the comparable statistics for the exact-answer scoring. In addition, there were fewer non-functioning items in the acceptable-answer versions.

Combining the hit-or-miss and modification methods, we note that Form AC C has the highest overall ID value at .38, an IF value very close to the ideal of .50 at .49, and only one non-functioning item. Thus so far, based on the descriptive statistics and item analyses, Form AC C appears to be the best overall from among the five original cloze tests scored for exact or acceptable answers.

As for the tailored-cloze method in the present study, the two cloze versions that were tailored proved to be far superior to any of the original versions within the same scoring schemes in terms of having better-centered mean IF values and higher mean item discrimination values (as a quick comparison of Tables 2 and Table 5 will reveal).

Based on the item analyses alone, we cannot definitively say which of the tailored cloze tests is likely to function the best in the future administrations. Revised EX appears to have the best-centered mean IF, and Revised AC clearly has the highest mean ID, while both tailored versions have no non-functioning items.
3. What Is the Effect on Reliability of Applying the Hit-or-Miss, Modification, and Tailored-Cloze Methods?

Finally examining the reliability estimates and the hit-or-miss method, the selection of which test works best from among the five exact-answer original cloze versions is relatively straightforward: EX B is the most reliable test according to the Cronbach alpha and split-half strategies (see Table 3). Since EX B also had the second highest mean ID and standard deviation reasonably good centering based on the mean and mean IF, EX B would appear to be a reasonably good candidate for selection as the best overall cloze test for exact-answer scoring if we were using the hit-or-miss method.

As for applying the modification method, we can again say that modifying the scoring scheme to the acceptable-answer scoring clearly had a salutary effect on the reliability of all the cloze tests in this study. As can be seen in Table 3, for all forms but one (Form B), the reliability estimates for acceptable-answer scoring are higher than the comparable statistics for the exact-answer scoring.

Combining the hit-or-miss and modification methods, we note that Form AC C has the highest Cronbach alpha and split-half (adjusted) reliability estimates at .832 and .814, respectively. Thus, using the hit-or-miss and modification methods, Form AC C appears to be the best overall from among the five original cloze versions scored for exact or acceptable answers.

As for the tailored-cloze method, the tailored-cloze versions proved to be superior to any of the original versions within the same scoring schemes in terms of reliability estimates. Based on the descriptive statistics, item analyses, and reliability estimates, we can only say for sure, at this point, that the acceptable-answer revision would likely function reasonably well in future administrations. While Revised AC is not the best centered, both Revised EX and Revised AC are reasonably well-centered as indicated by the means and mean IFs and neither has any non-functioning items. At the same time, Revised AC has the highest standard deviations, highest mean IDs, highest ranges, and the highest reliability estimates.

CONCLUSION

To summarize a bit, using the hit-or-miss method alone, we concluded that EX B was the single most reliable cloze test from among the five original versions because it had the highest Cronbach alpha and split-half reliability estimates, at .757 and .803,
respectively. Since this method also had the second highest mean ID (.26) and second highest standard deviation (3.76), we selected it even though it was third best in terms of centering based on the mean (9.33) and mean IF (.31), and even though it had nine non-fitting items. Other teachers might make different choices depending on which testing characteristics are most important to them.

Using the modification method, we found that modifying the scoring to the acceptable-answer scoring from the exact-answer scoring clearly had an overall salutary effect on the descriptive statistics, item analyses, and reliability estimates of the cloze tests in this study with only one exception (Form B produced reliability estimates that did not fit this pattern).

Combining the hit-or-miss and modification methods, we concluded that AC C had the highest Cronbach alpha and split-half (adjusted) reliability estimates at .832 and .814, respectively. Since that form was also reasonably well centered as indicated by the mean (14.76) and mean IF (.49) and had the highest standard deviation (6.29), as well as reasonably high range (25) and small number of non-functioning items (1), Form AC C appears to be the best overall cloze test from among the five original versions scored for exact or acceptable answers. However, once again, other teachers might make different choices depending on which testing characteristics are most important to them.

Finally, using the tailored-cloze method proved generally beneficial. The two cloze tests shown in Tables 4, 5, and 6 were both reasonably well centered as indicated by the means and mean IFs, and both had higher standard deviations and comparable or higher mean IDs than the five original versions. In addition, the revised tests had no non-functioning items. Between the two revised cloze tests, Revised AC had the highest standard deviation (6.29), highest mean ID (.47), highest range (25), and highest reliability estimate (alpha = .869). In short, if we were going to bet on which single test and scoring scheme would work best for norm-referenced purposes in the population studied here, the acceptable-answer revision (Revised AC) would be where we would put our money.

At this point, we must caution readers that all of these relationships might have changed if the levels of ability in the groups of students had been higher. For instance, if the abilities had been say seven points higher in terms of scores, all of the means for all the cloze forms would have been above the midpoint of 15 (half way between 0 and 30). Under those conditions, the means for the acceptable-answer scoring might have been
more poorly centered than those for the exact-answer scoring and the acceptable-answer standard deviations, ranges, and reliability estimates might all have been lower than the ones for the exact-answer scoring. However, within Japan, it is not likely that there are many groups of students who would be that much higher than the students we tested. Therefore, in most testing situations in Japan, we believe that similar patterns would be obtained if this study were replicated. We also assume that using the combination of hit-and-miss, modification, and tailored-cloze methods shown in this paper, anyone would once again be able to isolate cloze tests that produce superior distributions (in terms of means, standard deviations, and ranges), better item statistics (in terms of item facility values, item discrimination indexes, and number of non-functioning items), and higher reliability estimates than would have occurred if they just took a passage off the shelf and turned it into a cloze test.
REFERENCES


APPENDIX A:
EXAMPLE CLOZE TEST

Name ___________________________ Your nationality ___________________________

(Last) (First)

How much time have you spent in English speaking countries?

Directions: Fill in one word in each blank. You may write directly on the test.
Example: The girl was walking down the street when she stepped on some ice and fell (ex. 1) down.

---

The Science of Automatic Control (Form E)

The science of automatic control depends on certain common principles by which an organism, machine, or system regulates itself. Many historical developments up to the present day have helped to identify these principles.

For hundreds of years (1) were many examples of automatic control systems, but no connections (2) recognized among them. A very early example was a device (3) windmills designed to keep their sails facing into the wind. (4) consisted simply of a miniature windmill which rotated the whole (5) to face in any direction. The small mill was (6) right angles to the main one, and whenever the latter faced (7) the wrong direction, the wind caught the small mill’s sails (8) rotated the main mill to the correct position. Other automatic (9) mechanisms were invented with the development of steam power: first (10) engine governor, and then the steering engine controller... (continues for a total of 30 items)
APPENDIX B:
EXAMPLE CLOZE ANSWER KEY

Form E - exact answers (capitalized) and additional acceptable answers (lower case)

1. THERE
2. WERE easily originally
3. ON with for
4. IT this
5. MILL windmill mechanism device apparatus contrivance thing system
6. AT
7. IN to
8. AND then which
9. CONTROL
10. THE an

... (continues for a total of 30 items)
TAILORING CLOZE: THREE WAYS TO IMPROVE CLOZE TESTS

J. D. Brown  
Department of ESL  
University of Hawai'i  
1890 East-West Road  
Honolulu, HI 96822

brownj@hawaii.edu

Amy D. Yamashiro  
Nihon University, College of Commerce  
Kinuta 5-2-1, Setagaya-ku  
Tokyo, Japan 157-0073

amy@gol.com

Ethel Ogane  
Meiji University  
Department of Politics and Economics  
1-1 Surugadai, Kanda  
Chiyoda Ward, Tokyo, Japan 101-8301

ethel@gol.com