Shiken: JALT Testing & Evaluation SIG Newsletter, 1 (1) April 1997 (p. 2 - 11)

A cloze look at placement testing[1]

Christopher Jon Poel and Spencer D. Weatherly

How to choose the most appropriate method of estimating learners' knowledge or ability is never easy, particularly where learning a second language is concerned. Numerous commercial tests are available which purport to be able to accurately assess learners' language proficiency. In addition, many of these tests are described as being useful for placement testing. However, commercial tests are rarely, if ever, sensitive enough to accurately place a specific group of learners into their appropriate classroom level. So the question remains: When an institution wants to achieve a relatively consistent grouping of learners for placement purposes, what is the best method of estimating learners' abilities? The answer may be to use cloze tests.

Cohen (1980) reports on studies which have found high correlations between cloze test scores and tests of reading comprehension, (grammatically correct) writing ability, listening ability, and speaking proficiency, suggesting that the cloze test is an excellent indicator of overall language proficiency. Additionally, Bachman (1982) shows that cloze tests not only reflect low-level skills like phrase processing, but also complex skills such as human language processing capacity. At the same time, the evidence has not been conclusive. Other researchers have found much lower correlations, suggesting that cloze tests may not be valid or reliable as measures of general language proficiency[2].

"the single most important variable in the effectiveness (reliability and validity) of cloze may be how well a given passage fits a given population."

These contradictory findings may be the result of a mismatch between a particular cloze test and the group of students it is used with, much in the same way that commercial tests often do a poor job, as mentioned above. Brown (1988a, p. 20) suggests that the single most important variable in the effectiveness (reliability and validity) of cloze may be how well a given passage fits a given population. In other words, it may be necessary to provide a reading passage that is chosen for the specific learner population in order to obtain reliable and useful results. Brown, furthermore, goes on to demonstrate how relatively simple test-item analyses, namely item facility and item discrimination, can be used to improve the reliability of cloze tests. The resulting tailored cloze test is a good overall indicator of the learners' general language proficiency because it is designed specifically for a particular population of students.

In this paper we will report on a placement test development project in which we used the techniques outlined in Brown (1988a) to tailor a cloze test to accurately place vocational school students.

The Study

The project was undertaken at a two-year vocational school in Chiba City, Japan. In previous years, the students had been grouped based on either their first-term English grade point average (GPA) or an in-house version of the STEP-Eiken test. The resulting placement by either of these methods was less than satisfactory, and the students had to be extensively regrouped the following term. The effect of the reshuffling was a disruption of the harmony in the classes, as well as a loss of face by students who found themselves in a lower group. As a result of these experiences, the English department decided to develop a placement test specifically tailored to the school's particular student population.

Setting

The students at the vocational school are provided with training for the tourism and hotel industries. In each school year, they attend class for approximately 32 weeks, with eighteen 80-minute lesson periods per week. Of the total, four periods are devoted to English: English conversation, English for tourism, STEP-Eiken test preparation, and English reading/translation.

The students are between the ages of 18 and 21. All are high school graduates, meaning that they have had at least six years of English education in junior and senior high school. Slightly less than half are female. A large majority of the students enter this school after being unsuccessful in passing the entrance examinations for universities, although a small percentage enter because of personal interest in the tourism or hotel industries.

The students in each year are randomly divided into three homeroom groups, and these groupings are used for most of their classes. However, the faculty felt it would be more beneficial for the students and less frustrating for the teachers if they were grouped by proficiency for the English classes. Originally this was done after the first term, and was based on the students' English grades for that term. That policy was somewhat successful in the sense that the student groupings, when based on their actual English grades, quite accurately reflected their English ability. However, adjustments and regrouping were invariably necessary for the following term. Furthermore, during the first term teachers were saddled with mixed groups resulting in a less than optimal teaching (and learning) environment.

In an effort to have a more accurate placement procedure, the faculty devised a placement test that was given to incoming students the day before classes began. This test was an in-house version of the STEP-Eiken test, using questions adapted from the Level 3 and Level 4 tests. Since the test had to be marked, and groupings completed, by the following day, it was imperative that the test be easy to score. This constraint led to a rather short (20-items) and simple test, with the consequence

being that it did not do its job very well. After the first term, the students still had to be extensively regrouped based on the grades from their English classes back to square one.

In response to this latest problem, the native English teaching staff was asked to develop a more accurate placement test. Since the time allotted for administering and scoring the placement test was extremely limited, we decided that a multiple-choice cloze test would be a practical option. Recent research has indicated that a relatively short m-c cloze test can be a viable substitute for a longer test in measuring language proficiency (Ikeguchi, 1995).

Procedures

The first step in the multiple-choice cloze (m-c cloze) test project was a series of classroom assignments in which currently-enrolled students, working in pairs, were asked to fill in several cloze reading passages. Four readings were chosen in varying degrees of difficulty. They were chosen based on what the students could handle; that is, if we thought the students could read, comprehend, discuss, or answer questions about the passage, it was considered appropriate. Any story with numerous dialogues, proper names, or numbers was eliminated from consideration. Finally, we judged whether or not it would be of interest to the students. A reading that we thought would be appealing to the vocational school students was considered more appropriate (see Appendix for an example of one of the reading passages used in this study).

Because the cloze procedure was new to most of the students (that is, the students could not be expected to perform well on it), we felt that a lengthy distance between blanks was necessary in order to make it doable. At the same time, however, the goal of this project was to construct a multiple-choice cloze placement test, which is a much easier test. Therefore, in the final version the items would have to be closer together for the test to have any value at all. This dilemma was solved by making two versions of the cloze exercises for each reading.

In constructing the cloze exercises, we left the first two sentences intact in each of the readings. Starting with the third sentence, we counted until reaching the 13th word, which was deleted. This pattern was followed until the last sentence, which was also left intact. In the second version, the counting commenced with the sixth word of the third sentence, with every 13th word deleted. Later when we constructed the m-c cloze for each passage, we combined the items from the two fill-in versions to arrive at tests with an alternating 6th / 7th word deletion ratio[3].

The eight fill-in cloze exercises (four reading passages with two versions each) were given to second-year students as an in-class assignment. The second-year students were chosen for two reasons. First, we felt that the assignments would be too difficult for the first-year students. Second, we planned to pilot the multiple-choice cloze tests on the first-year students at the end of the year.

The second-year students worked in pairs, thus making it more likely that they would fill in all of the blanks and providing the largest possible pool of answers from which to choose distracters, as explained below. They were strongly encouraged to try to fill in all the blanks, and in fact did manage to complete the assignment in a majority of cases.

After the students completed the fill-in-the-blank cloze exercises, we wrote down all of the incorrect responses for each question along with the number of occurrences for each incorrect response. We then examined these incorrect answers and chose three as distracters for the m-c cloze pilot tests. In general, the three most often occurring incorrect answers were chosen. However, in some instances the answers were chosen by grammatical category. For example, if the correct answer was a preposition, all prepositions were chosen versus mixing prepositions with other grammatical forms such as nouns or verbs.

When this was done, the four choices for each item were then randomized. To get a truly random assignment, it is necessary to use a system that will not result in the test-maker choosing the placement of the choices off the top of her or his head, which could result in an unintentional pattern for the correct answers. The simple procedure that was used in this study involves using four playing cards: ace, two, three, and four. Shuffle the cards and then draw one: the number on the card indicates its position in the multiple-choice grouping. For example, for the first choice, watching, a playing card was drawn. It was a 3, therefore watching took the third position and received the letter designation c. This was repeated for each item on the whole test.

After the distracters had been chosen and randomized, the pilot version was administered to the first-year students at the end of the school year. This group was chosen because the students were most comparable to the student body that needed to be placed in the coming year.

After we piloted the multiple-choice tests, we scored each item dichotomously, that is either correct (0) or incorrect (1). The scores were then analyzed to determine which items were not contributing much, if anything, to the overall reliability of the test. Those items were then eliminated. Two standard statistics were used (for further discussion see Brown, 1988a, b, 1996):

* Item Facility (IF) the percentage of students who got the item correct.
* Item Discrimination (ID) a measure of how well the item separates the high-scoring students from the low-scoring students

After calculating the IF and the ID values, we found that the pilot test had an IF range of .075 to .830 and an ID range of .250 to .542. A range of .3 to .7 (that is, between 30% and 70% of students answer correctly) for the IF is considered to be desirable (Brown, 1996, p. 70). Items that fall within this range are at a level most appropriate for the specific learner population, as indicated

by the percentage of correct answers. Items that do not fall within this range are, in general, considered to be too easy or too difficult. We therefore chose to use this range as the standard for our item analysis.

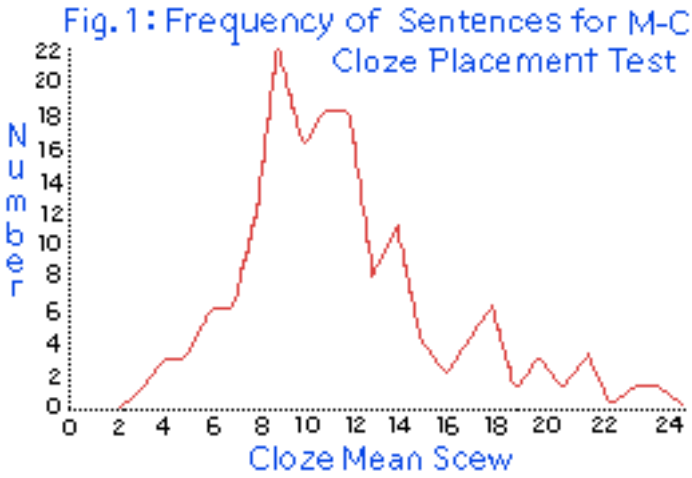
For this type of test the higher the ID the better. A high ID indicates that the item separates the more-proficient learners from the less-proficient ones. In general, items that have IFs within the range mentioned above will have fairly high IDs such that one criterion does not conflict with the other in the process of choosing items.

An example of a good item from our pilot test is one with an IF of .642 and an ID of .479. If we could write tests in which all of the items looked like this, we would probably be rich and famous or at the very least not have to work at a vocational school. An example of a bad item from the pilot test is one with an IF of .132 and ID of .083, which indicates that very few students answered the item correctly, and it is not separating the better students from the poorer ones for all practical purposes. One final example is an item that is both good and bad. It has an IF of .377, which is within the parameters we set, and at first looks like a keeper. The ID on the other hand is .021 which tells us that, in fact, the low-level students are getting the item correct and the high-level ones are not, as indicated by the negative ID. Therefore, even though the item has a preferable IF, the ID is clearly unacceptable and so the item was discarded. This clearly demonstrates the necessity of calculating both IF and ID when doing item analysis. After the item analysis was completed, we selected the two passages with the most reliable items to create our final version. The result was a 31-item multiple-choice cloze test comprised of two reading passages.

The resulting 31-item multiple-choice cloze test was administered immediately before the start of the following school year, along with the usual 20-item in-house test, to incoming students. The tests were then scored and students were ranked according to their total scores our test plus the in-house test. From these ranks, the students were placed into six groups. At the end of the first term, the students' results in the four English courses were calculated and used for correlation analysis.

Results and Discussion

This test had a mean of 11.440 and a median of 11, showing that it was fairly well balanced, but slightly off-center. This can be seen clearly in the graph in Figure 1, where the slight positive skew is shown.



Results of the correlation analysis are shown in Table 1. As can be seen, the cloze test correlates quite highly with the students' actual rank (r =.797). This is to be expected, because in large part the test is correlating with itself, since it was partially responsible for the placement decisions. More to the point of this paper, however, are the correlations that were observed between the placement rankings based on the cloze test and the students' English grade-point average (GPA) rankings (r =.509). While this correlation appears to be low, it does show an adequate degree of relationship between the cloze test and the students' actual performance in the classroom. In addition, it is certainly better than the traditional method using the in-house STEP-Eiken style test (r =.350).

Table 1. *Matrix of Spearman correlation coefficients*

	Actual	Cloze	In-House	Combined	GPA
Actual Rank	1.00				
Cloze Rank Test	0.797	1.000			
In-House Rank Test	0.758	0.354	1.000		
Combined Test Rank	0.959	0.811	0.765	1.000	
GPA Rank	0.495	0.509	0.350	0.525	1.000

Number of Observations: 150

One major reason for the low correlations observed is the nature of the rankings used in this study. Rather than individually ranking each student, i.e., from 1 to 150, we ranked the students based on their group designation as a result of the test, from 1 to 6. This had the effect of collapsing the data, which resulted in numerous fuzzy cases where students on the border of two groups had the same score. Thus, the Spearman rank-order correlations that we calculated were based on a range of 1 to 6. This clearly contributed to the low correlations that were found. Likewise, when we estimated groupings based on GPA, students with the same average were randomly assigned to one of the two groups, which clouded the issue further. In essence, then, we correlated the students actual group (determined by our placement test) with their estimated group (based on their GPA), with both rankings having a severely limited range of 1 to 6. Under such circumstances, a correlation of r = .509 is remarkably high.

Two additional statistics were calculated for the multiple-choice cloze test: the reliability and the standard error of measurement (SEM) in Table 2. Reliability ranges from r = .603 (KR-21) to r

= 669 (Cronbach alpha). Although this is not as high as we would have liked, these scores are adequate considering the length of the test involved. Estimating the reliability for a 50-item test using the Spearman-Brown prophesy formula (Brown, 1996, p. 195), results in a reliability range of r =.710 to r =.765, which, while still quite low, is much better.

Table 2. *Reliability and SEM for a multiple-choice cloze placement test*

	Reliability	SEM	Estimated r for 50-item test
Cronbach Alpha	0.699	2.397	0.765
KR-20	0.656	2.433	0.754
KR-21	0.603	2.623	0.710

Of more concern for the usefulness of the test is the standard error of measurement. As seen in Table 3, the SEM was approximately 2.5. This means that a student who scored 15 on the cloze test could be expected to score between 12.5 and 17.5 about 68% of the time if the test were given repeatedly. Given that we need to make very fine decisions with this placement test (as discussed below), an SEM this high is unacceptable. Because the SEM is directly related to reliability, increasing the reliability would result in a lower SEM and a more accurate test.

Given, then, that the multiple-choice cloze placement test created for this study was not perfect, the question arises: What can be done to improve the procedure? We consider three steps to be important in developing a more effective placement test.

First, a more careful selection of reading passages must be carried out. It is especially important to remember that the passages should be chosen with regard to the students' English productive proficiency level rather than receptive proficiency. The cloze procedure that was used in gathering distracters, as explained above, is a productive test. Students have to fill in the blanks without any choices or hints other than the surrounding text. If the reading passage chosen is one which the students can normally read and comprehend (receptive mode) but which they cannot discuss to any suitable degree (productive mode), it is probably not useful to use this procedure. The students, when confronted by such a passage, are often reduced to blind guessing and are not able to provide good distracters. If, however, the passage is at the students' productive level of proficiency, the procedures outlined in this paper will be valuable in providing the necessary distracters, allowing for the development of a more meaningful test.

This problem became apparent in the distracter-generating phase of our study, when we discovered that the number of students who provided incorrect answers, and thus potential distracters, was quite low for some items. As explained above, the students had been given several cloze passages to work on in pairs, and they were strongly encouraged to try to fill in every blank.

However, this was the first time that most of them had ever seen a cloze passage, and many of them were confused as to what to actually do, resulting in lost time needed to repeatedly explain the procedure.

This was a major difficulty when it came time to tally up the incorrect replies that were to be used for distracters. For many of the original blanks, the choice of distracters was limited (or, in a few cases, nonexistent), and the item had to be thrown out even before starting. At other times, we were forced to go ahead with distracters that had been provided by an extremely small number of students in more than one case, the third distracter was one which had been given by a single student.

An additional problem that appeared later when constructing the multiple-choice tests was that the reading passages were too short. On the final placement test, we went ahead with two passages which had 15 and 16 items, respectively. While 31 items is considered adequate by most testing experts, we feel that it was too few considering that 150 students were to be placed into only six groups on the basis of these scores. At the time we did not consider this to be such a serious problem because of the inclusion of the 20-item in-house test made by the Japanese staff. However, as can be seen in Table 1, the in-house test added little to the strength of the correlation (r=.509 to r=.525). Thus it may be that increasing the number of cloze-test items will improve the accuracy and reliability of the placements.

Finally, while the topic of the two readings used in the final test were of interest to the students, their appropriateness is questionable. The two that were ultimately used were both about American culture, one dealing with Americans television viewing habits (see Appendix), and the other with American vacation patterns. It would be better to provide reading passages which were more culturally relevant for the students so that, instead of measuring their (American) cultural proficiency, we would be measuring their English language proficiency more directly.

Therefore, in the next phase of the placement test development project, we will attempt to choose reading passages with closer cultural relevancy for the Japanese young adult population we are dealing with. Additional improvements in the procedure will include having the students work on the original cloze passages individually rather than in pairs, thus providing more incorrect replies and making our distracter selection much more solid. Finally, the passages will be made longer, from an average of 225 words to 300 words. Although increasing the length introduces the potential for test fatigue, this should not be a problem as the revised test will still be only approximately 600 words long. Given that the time allotted for the test is 60 minutes, it is not considered too long to constitute a problem. And, of course, the longer passages will make it possible for us to construct two 25-item multiple-choice cloze tests for the desired length of 50 items total.

Conclusion

While the multiple-choice cloze placement test developed in this project was less effective than it might have been, it provides an important foundation on which to build. Basic item analysis statistics – item facility and item discrimination – have proven to be easy to apply and effective in constructing test items. By eliminating the observed procedural errors discussed above, an improved multiple-choice cloze placement test can be developed. It is hoped that the new test will provide a solution to the problem of placing our students.

Notes

1. An earlier version of this paper was presented by the authors at the 16th Annual JALT Conference in Omiya, Japan in November 1990.
2. See Ikeguchi (1995) for an excellent discussion of this and other issues pertaining to cloze testing.
3. Brown (1983, 1988a) suggests that cloze passages can best be made to fit a particular group when the distance between items is no less than five words, and no more than nine. How or why we ended up with a 6th/7th deletion pattern is a mystery to us.

References

Bachman, L. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16* (1) 61-70.

Brown, J.D. (1983). A closer look at cloze: Validity and reliability. In J.W. Oller, Jr. (Ed.), *Issues in Language Testing*. Cambridge, MA: Newbury House.

Brown, J.D. (1988a). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5* (1) 19-31.

Brown, J.D. (1988b). *Understanding Research in Second Language Learning: A Teachers Guide to Statistics and Research Design*. Cambridge: Cambridge University Press.

Brown, J.D. (1996). *Testing in Language Programs*. Englewood Cliffs, NJ: Prentice Hall Regents.

Cohen, A.C. (1980). *Testing Language Ability in the Classroom*. Rowley, MA: Newbury House.

Decker, R. (1982). *Windows on the U.S.A.* (500 word level). Chicago, IL: Science Research Associates.

Ikeguchi, C. (1995). Cloze testing options for the classroom. In J.D. Brown and S. Okada Yamashita (Eds.), *Language Testing In Japan,* p. 166-178. Tokyo: JALT.

Appendix: A Sample Cloze Test

American boys and girls love to watch television. Some children spend six hours a day in school and four to six hours a day in front of the television set. Some children even (1) _____ television for eight hours on (2) _____ or Saturday. But many parents let their (3) _____ watch only during certain hours.

Television shows are like books or movies. A child can learn bad things (4) _____ some of them and good things from others. Some shows help children understand the news from Washington and (5) _____ parts of the world. Some (6) _____ show people and places from other countries or other times in history. With television a child does not have (7) _____ go to the zoo to (8) _____ animals or to the ocean to (9) _____ a ship. Boys and girls (10) _____ watch a play, a concert, or a baseball game at home. Some programs even teach children how to cook (11) _____ how to use tools.

Television (12) _____ many places and events into the living rooms of our homes. Some (13) _____ show crime and other things that (14) _____ bad for children, so parents (15) _____ help them to find other activities that are interesting.

It is fun to watch television, but it is also fun to play a musical instrument, to read a book, or to visit with friends. It is important for children to have many different things that they are interested in.

```
(1) _____
        (A) watches  (B) watch (C) watching (D) watch
(2) _____
        (A) Friday  (B) more (C) less (D) ten
(3) _____
        (A) are  (B) children (C) with (D) television
(4) _____
        (A) to  (B) from (C) by (D) for
(5) _____
        (A) other  (B) New York (C) almost (D) American
(6) _____
        (A) parents  (B) children (C) programs (D) television
(7) _____
        (A) to  (B) had (C) been (D) time
(8) _____
        (A) visit (B) see (C) an (D) look
(9) _____
        (A) ride (B) look (C) see (D) get
(10) _____
        (A) want (B) television (C) can (D) can't
(11) _____
        (A) do (B) also (C) of (D) are
(12) _____
        (A) is  (B) how (C) shows (D) brings
(13) _____
        (A) programs (B) parents (C) television (D) places
(14) _____
        (A) are (B) not (C) to (D) very
(15) _____
        (A) don't (B) never (C) to (D) sometimes
```