JALT Hokkaido Journal Vol. 7 pp. 24-34 journal@jalthokkaido.org

The Effectiveness of Cloze Testing for Japanese Learners of English

Mark Chapman

mark@ilcs.hokudai.ac.jp

Hokkaido University

Abstract

This paper attempts to outline some difficulties in the application of cloze testing to Japanese learners of English. Despite cloze testing being an apparently simple and efficient way of measuring language proficiency, there are a number of problematic areas involved in cloze test construction and scoring. This paper focuses on two main areas; firstly a summary of a recent lecture given by J. D. Brown on problems associated with cloze testing and secondly an independent study that examined the construction and scoring of a cloze test for a group of Japanese subjects. The main problems highlighted are the lengthy trialing required if the cloze test is to be reliable and the difficulties involved in selecting an appropriate scoring procedure.

日本人英語学習者に対するクローズテストの有効性

この文書では英語を学ぶ日本人にクローズテストを適用する際に生じるいくつかの問題点の概要を述べることを試みる。クローズテストは一見簡単で、言語の上達具合をはかる上で能率的な方法であるにもかかわらず、その構造と採点法の分野を含む数多くの問題領域がある。ここでは二つの主な領域に焦点をあてることにする。まずはじめにクローズテストに関する問題について行われた J　D　Brown 氏の講義について、そして二つ目にはある日本人グループを対象に行われたクローズテストの構造と採点法を調査した独自の研究についてである。強調されるべき主な問題は、クローズテストを信頼性のあるものにするためには長期の試用が必要なことと、適当な採点手続きを選ぶ際の難しさについてである。

Introduction

Cloze testing is an appealing way of testing students' overall language proficiency. The tests are quick and easy to construct and can be marked with minimal effort if the most simplistic scoring method is employed. Especially if utilized as part of a university entrance exam or as a placement test, cloze tests would seem to be an efficient and reliable way of gaining a snapshot of learners' ability. These tests also have face validity with learners; cloze tests look like a 'real test.' However, in a recent lecture given by J. D. Brown in Japan, doubt was cast over the usefulness of cloze testing for the average Japanese learner of English. Professor Brown is a leading expert in the field of language testing from the University of Hawaii and is the author of several publications related to cloze testing. This paper will give a brief summary of Brown's findings and will also report on an independent study into cloze test construction and scoring.

What is a cloze test?

A traditional cloze test takes a text and deletes a certain number of words from it. The task of the test taker is to replace the missing words. The traditional cloze test has a mechanical deletion rate whereby every *nth* word is deleted. The alternative is a selective deletion exercise whereby the deleted words are not selected mechanically but rather are deleted at the discretion of the test maker. A discussion of the advantages and disadvantages of the two methods will follow later.

What do cloze tests measure?

It has been suggested that cloze tests provide a reasonably accurate measure of overall linguistic ability. The cloze test may at first appear to be simply a test of reading ability but this has been questioned. Weir (1993:79) suggests that a cloze test tests a more limited range of reading skills than do short-answer questions. Weir questions cloze tests' capacity to measure the ability to skim a text or to read it carefully to understand its main ideas and details. Weir does support cloze as a measure of general proficiency however:

> The format may well have a place, though, if one is looking for a quick, efficient measure of general language proficiency for the purpose of placement of students into language classes.

This view is supported by Hughes (1989) who also sees cloze as a satisfactory test of general language proficiency provided that “decisions which might have serious adverse effects” are not taken solely on the basis of the results of the cloze test.

Summary of Brown’s remarks

In a lecture given in May 2003 in Kyoto Brown raised several problems with cloze tests. A clear distinction was made between the effectiveness of natural and tailored cloze tests. A natural cloze test takes an authentic text at random and employs mechanical deletion. The appeal of natural cloze tests lies in the lack of subjectivity in the creation of the test and its apparent fairness. Brown (1993) found that natural cloze tests produce consistently low scores and fail to discriminate among candidates. In an investigation that included fifty separate natural cloze tests administered to groups of between 40 and 50 subjects, no group averaged more than 10 points out of 30 and only 14 of the 50 groups averaged more than five points out of 30. These tests were not administered to particularly low-level learners, but non-native speaking students at UCLA. Given the size of this study (n=2000+) it is likely that the subjects represent a heterogeneous group. Despite the very low scores the cloze tests produced consistently high reliability coefficients, which suggests the tests are good ones. However, these high coefficients demonstrate that natural cloze tests generally fail to discriminate between candidates.

Brown suggests that tailored cloze tests are considerably more effective. A tailored cloze test requires trialing with a large group of learners similar in ability to the target population. After administering the test to the trial group the results must then be analyzed for item facility (which items are at the appropriate level of difficulty) and item discrimination (which items can identify differences in ability between members of the target population). After selecting approximately 30 items that meet a suitably high standard of item facility and item discrimination, the test can be reconstructed and administered to the target population. Brown reported significant problems in constructing reliable tailored cloze tests for low proficiency learners of English. In 2002 Brown performed an extensive study into tailored cloze tests at Keio University and Heisei University. Five versions of a cloze test (30 items on each test) were administered to 193 low proficiency students and 143 high proficiency learners. Unfortunately Brown did not go into detail on how candidates were assessed, but did comment that a large proportion of the subjects from Keio were returnees. The low proficiency subjects got very low mean scores on all versions, with the highest being approximately 4 points out of 30 by SEMAC (semantically acceptable) scoring. On the same test, the mean of the high proficiency group was approximately 15. Brown also commented on the following, troubling data. Very few items (10 from 150) were at an appropriate level of difficulty for the low proficiency learners. In contrast, 91 items from 150 were at an appropriate level of difficulty for the high proficiency group. The data was only slightly more encouraging for item discrimination. Less than 30 of the 150 items were able to discriminate among the low proficiency subjects, whereas 80 of the items had an acceptable level of item discrimination for the high proficiency group. Brown's data indicates that it will be very difficult to construct a reliable cloze test for low proficiency learners.

If cloze testing is to be effective in terms of reliability, then the tests must be trialed and carefully constructed. The test designers will also need a basic working knowledge of statistics including how to measure item facility and item discrimination. Even given these requirements, it is still doubtful whether the test will be a valid one for low proficiency learners. Brown admitted that his high proficiency group at Keio University included a large number of returnees and this fact makes the data even more troubling. The low proficiency group at Heisei may not be typical of all Japanese university students but they are likely to have a similar level to an average Japanese senior high school student. This suggests that cloze testing is unsuitable for university entrance exams. It also suggests that cloze tests are not a reliable placement test for English classes at Japanese universities with the exception of schools that have a generally high level of proficiency. It would seem that this precludes the majority of Japanese universities.

An independent study into cloze test scoring

This section will report on a recent study into the problems of constructing and scoring a cloze test for a group of intermediate learners of English in Japan. The purpose of the study was to attempt to verify existing claims made about scoring procedures for cloze testing. Testing literature suggests that the two methods of scoring cloze tests (see below) correlate highly and do not significantly alter the ranking of students. If the two systems do correlate at an acceptably high level then cloze tests can be reliably scored by the faster and more efficient method.

Method

Participants

The test was administered to a group of twenty-six students who had enrolled in a three-week intensive English course for employees of Hitachi Ltd. in Japan. The students were of an intermediate level with TOEIC scores between 450 and 650. During the course the students had read two guided readers and had moved on to read three short unedited newspaper articles in preparation for discussion classes. They had also read a number of edited newspaper articles for the purpose of speed reading classes.

Selecting a text

The participants were almost exclusively from scientific backgrounds and this influenced the topic selected for the text. An investigation by Alderson and Urquhart (1983) suggested that science students find texts outside the scientific field very challenging, more so than for students of other subjects tackling a science-related text. This reinforced the belief that the text should be on a science-related topic that would be familiar to as many of the test takers as possible. The text finally selected was about internet providers (see appendix 1). Some Hitachi employees are now actively involved in research fields closely related to the Internet so it is possible that such students would have an unfair advantage.

Mutilating the text

The choice here is whether to opt for a traditional cloze exercise with a mechanical deletion rate or to utilize a selective deletion approach. An argument in favour of mechanical deletion is that it is simple and easy to construct. If the deletion rate is relatively short, say every fifth word, then a large number of items can be tested with a relatively short text. It has also been suggested (Hughes 1989) that a mechanical deletion rate can allow a representative sample of linguistic features of the text to be deleted. Mechanical deletion also removes an element of subjectivity from the choice of which words to delete. Once the first deletion has been made, the test maker cannot influence which words will be removed.

The advantage of mechanical deletion being less subjective than selective deletion also raises a problem. With a mechanical deletion rate there will almost always be problematic items raised. Bailey (1998) reports an incident where one of the items deleted from the text through mechanical deletion was the word *nacre*, which none of the native speakers involved in the creation of the test knew the meaning of. Items that are impossible to replace are irritating to the test takers and may damage the face validity of the test.

An advantage of selective deletion is that the deleted items are easy to modify if they are shown to be problematic during pre-testing. Replacing problematic items in a mechanically deleted text entails changing all the items. A further advantage reported by Alderson (1995) is that selective deletion enables the test taker to focus on the aspects of language that are of interest. This allows the deleted items to be in line with the test construct.

After considering both forms of deletion it was decided to proceed with selective deletion. The main reason was to avoid problematic items, especially as the text chosen contained a considerable number of names, dates, numbers and facts that would have been impossible to replace. Both lexical and grammatical items were deleted in order to provide as thorough a test as possible of the participants' general proficiency. A copy of test can be found at appendix 1.

It was also decided to leave a lead-in to the text unmutilated with no deletions made in the first paragraph. This approach is recommended by Hughes (1989) and Weir (1990). An unmutilated first paragraph allows the reader to find his or her feet as it were and gain a basic understanding of the subject of the text, without which it may be excessively difficult to begin the replacement of the deleted items.

The final area of concern in constructing the test was the instructions given to the subjects. It is essential for the instructions to be clear and simple. Alderson (1995) emphasizes that the subjects must be unambiguously told whether each gap should be filled with one word or two. In the case of this investigation only one word was allowed so contractions such as *we'll* were avoided, as were hyphenated words.

Administering the test

The participants were given one hour to complete all the deleted items under controlled conditions in that they were not allowed to confer with each other. They were however

scoring due to its simplicity or convenience. There would still be one participant who ranked seven positions higher on SEMAC (moving from a mediocre 10th place to a creditable 3rd) and one participant who ranked six places higher on SEMAC. These two cases may not be statistically significant but they are very significant to the two candidates themselves.

The results reported here do not justify the claim that SEMAC does not change the ranking of subjects on a cloze test in comparison with exact scores. From the twenty-six participants who took the test, ten of them or 38.5% had their ranking altered by four or more places. Only eight of the participants had their rankings unaltered by SEMAC scoring. Four more participants had their ranking altered by only one position, making a total of twelve subjects or 46% who had rankings that were minimally changed.

Despite the small scale of this study and the lack of objective criteria for deciding what constitutes a semantically acceptable response, the data does seem to indicate that there is room for doubt over the worthiness of exact word scoring. Clearly, exact word scoring offers advantages in terms of speed of marking and reliability. However, these practical advantages would not seem to offset the problem of the discrepancies in ranking illustrated in this study. SEMAC scoring with objective criteria for specifying what is an acceptable response (preferably with a trial group of native speakers) would provide a more valid test.

Conclusion

Cloze testing is certainly not a simple means of getting a quick view of learners' general language proficiency as once thought. If cloze testing is to be valid and reliable, the test must be carefully constructed (see appendix 2 for specific guidance). Ideally, the cloze will be trialed with learners similar to the target group to select appropriate items for deletion. If the test is to be marked by SEMAC scoring, the acceptable answers should be defined by trialing the test on a large number of native speakers. SEMAC scoring will almost always be preferable to exact word scoring unless the test score has little significant impact on the subjects. If cloze tests are constructed and marked in accordance with these guidelines, they still have a meaningful role to play in language testing for Japanese learners. However, test makers must be careful not to abuse the apparent ease of construction as the price to pay for a hurried construction is a test that provides at best questionable information about the subjects.

References

Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C. and Urquhart, A. H. (1983). *The effect of student background discipline on comprehension: a pilot study.* From Hughes, A. (1983). *Current developments in language testing* (pp. 63-74). London: Academic Press Inc.

Bailey, K. M. (1998) *Learning about language assessment: Dilemmas, decisions and directions.* Boston, MA: Heinle and Heinle.

Brown, J. D. (1993). What are the characteristics of *natural* cloze tests? *Language Testing, 10,* 93-116.

Brown, J. D. (2002). Do cloze tests work, or is it just an illusion? *Second Language Studies: Working Papers of the Department of Second Language Studies*, *20*, University of Hawaii.

Brown, J. D. (2003). Norm-referenced item analysis (item facility and item discrimination). *Shiken*, *7*, 16-19.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Owen, C. (1997). *Testing.* University of Birmingham.

Weir, C. J. (1990). *Communicative language testing*. Exeter: University of Exeter: Prentice Hall.

Weir, C. J. (1993). *Understanding and developing language tests*. New York: Prentice Hall.

Appendix 1

There are 25 words missing from the following article. Please choose the most suitable word (only one word) for each gap.

Internet-connection services based on communication satellites, which offer much faster data downloads than are available via phone lines have started picking up customers in Japan. “People feel satellite-based connections offer as much speed as connections through leased lines, or are even faster” said Yoshiyuki Ito, marketing manager (1) Direct Internet Corp. Satellite based internet connections reach data speeds of 400 kilobits (2) second, while connections through integrated service digital network (ISDN) phone lines are regarded as (3) at 128 kilobits per second. Direct Internet, a leading provider of satellite-based internet connections in Japan, (4) offering the service in August 1997 using PanAmSat Corp.’s PAS-2 communications satellite, which orbits (5) Hawaii.

Direct Internet’s service is not regarded as suitable for personal (6) because a customer needs to pay around 150,000 Yen (7) buy a 60 cm-diameter parabolic antenna and other equipment. In addition, each customer needs internet (8) via a regular phone or leased line, because the satellite link can be (9) only for downloading. Direct Internet has (10) targeting businesses as its customers since its establishment (11) November 7th 1996. This past February, the company started building an information network for Autobacs Seven Co.’s autoparts stores in Japan and overseas. The system, to be up and (12) in October, will distribute product information via satellite.

Direct Internet has also received an order to build a satellite-based information system for the operator of a major (13) of sporting-goods shops. The system will gather point-of-sale data from shops via phone lines, and send information on products that are (14) well via the PAS-2 satellite. “Satellite-based internet connections are especially useful (15) chain-shop operators,” said Hiroshi Kubori, Direct Internet’s managing director. For example, product promotions can be carried (16) more effectively by having promotional images that can be easily changed. Some convenience stores (17) product-promotion videos in their shops, but they normally play the (18) video over and over.

Starting this month, Direct Internet (19) offer a satellite-based extranet service for small businesses. Extranets are internet-based networks (20) can be accessed by both internal and outside users. The service is expected to be helpful in situations where multiple companies need (21) share large-volume information, such as computer-aided design data, and collaborate online.

Direct Internet hopes to (22) its number of customers from 15 companies now to 100 in three years. The increasing number of small and home offices and the rising (23) for collaborative production of digital content are expected to expand its market. But competition is also likely to (24) tougher, as NTT entered the market through a (25) venture established in January with Japan Satellite Systems Corp. Kubori said Direct Internet aims to survive by focusing on providing good service.

Appendix 2

Professor Brown gave the following outline for constructing a reliable cloze test:

Create a large pool or around 150 items by finding a passage of appropriate length and topic for the average student in the population. Create 30 blanks at every *nth* word interval with *n* determined by the amount of context the students will need and the length of the passage. Then create four additional forms with different starting points for a total of five 30 item forms including 150 items.

Pilot the items by randomly distributing them in a fairly large group of students similar to the target population.

Perform item analysis to estimate the item facility and item discrimination (see Brown 2003; soon to be available online at the JALT TEVAL SIG website.) In each set of five items eliminate items that discriminate poorly (below ID=0.3), items that are outside the acceptable level of facility (IF 0.3 – 0.7) Remake the cloze passage with blanks only for items that are working well and discriminate well for the particular group in question.