How Close Is C-Test to Cloze?

Gary Simon

Introduction

Random deletion cloze procedure and the more recently developed C-Test are commonly used norm-referenced assessment instruments that compare and rank language learners. Both have their origins in *Information Theory* which is based on research that found normal human communication has multiple built-in levels of redundancy (Miller, 1969) (Klein-Braley 1985:33-35) (Oller 1978:56). This redundancy ensures messages are understood despite various kinds of interference or "noise." Although both of these *reduced redundancy tests* (Spolsky, 1969) have a great deal in common theoretically, differences in construction raise questions as to which is a better means of assessing learners abilities.

Random deletion cloze procedure has a relatively long history. It was first developed by Wilson Taylor who proposed that a measure of readability could be derived by systematically deleting words from a text and having native speakers fill in the blanks (Taylor, 1953). Subsequent research led him to conclude that it could also be used as a reading comprehension test (Taylor, 1957).

In terms of ESL/EFL assessment, cloze procedure has generally

come to mean a test passage of between 250 and 400 words (Silberstein 1989:32) (Madsen 1983:48). The first and final sentences of the chosen text are left complete to provide a contextual frame. Then, beginning somewhere in the second sentence, words are deleted at regular intervals. Proper nouns and numbers are not deleted. Depending on what pattern of deletion is chosen, every 5th, 6th, 7th, 8th, 9th, or 10th word is omitted from the passage. Such "nth," "fixed-ratio," or "random" deletion patterns are believed to give an objective sampling of the words in the text and can yield approximately 50 items to be analyzed (Valette 1977:212). Over the years, two methods of scoring have been used. One is *exact* scoring whereby the test taker is only given credit for a completion if he or she has supplied the same word that appeared in the original text. The second method is the *acceptable* means of scoring in which any close synonym, as well as the original word, is awarded a point. In both methods of scoring, however, the answer must be completely correct according to native-speaker standards of meaning, grammar, spelling and collocation (Brown 1983:111).

Proponents of random deletion cloze testing for nonnative learners of English, like John W. Oller, Jr., have referred to such tests as "pragmatic" (1978). It is believed that clozes examine the non-native speaker's ability to perform the same kinds of tasks as native speakers would in daily life. Oller has claimed that such tests are preferable to "...the discrete-point type in that they tap the underlying, internalized expectancy grammar of the examinee." (Oller 1978:55-56). As the test-taker fills in blanks, he or she is hypothesizing about the

meaning of a passage (Oller 1978:44). Oller, and many others, feel that such tests can better indicate a learner's level of proficiency because both receptive and productive language skills are being utilized in tandem. Since cloze procedure does not require "after-the-fact" questions whose wording may lead the test taker to respond in a specific direction, it is thought to be closer to the communicative process of spontaneous language exchange (Silberstein 1989:34).

Many researchers assert that random deletion cloze testing is a means of assessing a learner's global knowledge of a language, and that data rendered by this integrated testing procedure only affords information about an examinee's overall knowledge of grammar and vocabulary. However, others believe that this procedure may also engage higher level skills such as those required in writing and reading. For example, in independent studies at universities in Japan and Korea, Sandra Fotos (1991) and Soyoung Lee (1996) found that scores on random deletion cloze tests correlate very well with results on placement essay exams. Similarly, writing more than 40 years after Taylor's work, James D. Brown has also expressed the view that cloze procedure is a valuable tool in assessing both readability of ESL/EFL texts and reading comprehension. Brown feels there is convincing research that suggests the skills needed in cloze procedure are very similar to those required in the reading process (1998: 9–10). Another researcher of note, Lyle Bachman, refers to cloze as an "enigma," but states this procedure can assess all reading competencies. However, Bachman expresses very strong reservations about the random deletion process. He suggests that clozes used to measure

reading abilities should make use of rational deletion, in which selected items are deliberately chosen, to ensure the testing of specific competencies (1984:14–15).

Random deletion is but one aspect of this procedure that has brought traditional ESL/EFL cloze testing into question. The fact that there have been no accepted standard pattern or starting point of deletion has made it difficult to establish the reliability and validity of such tests. It has been found the same passage will yield a range of tests of different difficulty depending upon where deletion begins, and what nth–pattern is employed. It is felt that this compromises test reliability (Klein–Braley 1985:43–45). Furthermore, it has been found that the matching of suitable passages with the learning environment and backgrounds of examinees is a variable that must be considered when writing such tests (Brown 1983:118). Therefore, a great deal of trial testing is necessary to find appropriate clozes to assess accurately the competencies of any given group of learners.

In an attempt to maintain the same theoretical framework and address these issues, Ulrich Raatz and Christine Klein–Braley (1981), designed and researched an alternative to the random cloze test. The result of this work, the "C–Test," is a standardized norm–referencing procedure (Raatz 1984:125).

A C–Test, as designed by Raatz and Klein–Braley, consists of four to six *short* complete passages that focus on different topics. In each passage, the first sentence is left intact. Then, the "rule of two" is applied. This means that beginning with the second word in the second sentence, every other word is "damaged". A damaged word is

one in which the second half has been deleted. Numbers, proper nouns, and single letter words like "I" or "a" are left undamaged, and the pattern is resumed immediately after the appearance of one of these. This pattern is continued until there are 20 to 25 half-deleted words. Then the passage is allowed to continue to its conclusion without further alteration. Raatz's description is not entirely clear, however, it appears that words with odd numbers of letters establish a pattern wherein one letter more than exactly half is given to the first targeted word to appear, and one letter less than exactly half to the next (see below, "the," "other," etc.) (Raatz 1985:50).

Although the following would not be an appropriate passage for selection, if the preceding paragraph were one of the 4 to 6 texts in a C-Test, it would look like:

A C-Test, as designed by Klein-Braley and Raatz, consists of 4 to 6 short complete passages that focus on different topics. In ea_____ passage, th_____ first sent_____ is le_____ intact. Th_____, the "rule of two" i_____ applied. Th_____ means th_____ beginning wi_____ the second wo_____ in th_____ second sent_____, every ot_____ word i_____ "damaged". A dama_____ word i_____ one th_____ has be_____ half del_____. Numbers, pro_____ nouns, and single letter words like "I" or "a" are left undamaged, and the pattern is resumed immediately after the appearance of one of these. This pattern is continued until there are 20 to 25 half-deleted words. Then the passage is allowed to continue to its conclusion without further alteration. Raatz's description is not entirely clear, however, it appears that words with odd numbers of letters establish a pattern wherein one letter more than exactly half is given to the first targeted word to appear, and one letter less than exactly half to the next (see below, "the, "other," etc.) (Raatz 1985:50).

Test-takers are instructed to "repair" or complete the words. Raatz states that five to seven minutes should be given for each passage in the C-Test (Raatz 1985:49-51). Scoring is a relatively quick and easy task as the grader reads through a limited number of sentences. Because completion of each damaged item is contingent on the examinee understanding the complete context in which it is embedded, individual damaged words are not considered independent items when analyzed. Rather, each passage is a "superitem" whose results are only considered in terms of the whole test consisting of all of the texts. Reliability is established through Cronbach's Alpha coefficient (Raatz 1985:49-55).

As Raatz points out, individual passages for C-Tests can be much shorter than those used for clozes (1985:49). For example, the above sample superitem consists of 165 words, and a better selected passage would have a sufficient amount of context and redundancy for the vast majority of native speakers to complete correctly. In contrast, a single cloze using a deletion pattern of every 7th word, for example, would require a passage more than twice as long to yield a sampling of 50 words. This means a test taker should be able to apply his or her knowledge of a language to several C-Test passages with various kinds of content in a shorter period of time. An analysis of his or her responses to different topics in context would give a more balanced and reliable assessment of his or her abilities.

As C-tests are norm oriented, Klein-Braley states that the mean difficulty of the superitem passages as a group should be 50% to allow for a good distribution of scores. Therefore, some passages

within a test can be relatively easier or more difficult than others (1984:98). However, as does Brown with regard to cloze passages, Klein–Braley and Raatz stress the importance of matching the superitem texts with the examinees in content and difficulty. Klein–Braley cautions against using passages that require specialized knowledge or that contain archaic or colloquial usage of language. Because of cultural content, she also states that literary texts or passages meant to be humorous should not be used (1984:97). In addition, Raatz writes that texts should be especially constructed or modified for nonnative speakers of a language (Raatz 1985:53).

Klein–Braley, Raatz, and other researchers have done extensive testing of the procedure with native speakers of various languages, and with nonnative speakers in both foreign and second language learning environments (Raatz 1984:124–139). A large body of research confirms that the procedure is a reliable and valid means of norm referencing a learner's *general* competency level. Their investigations have found that C–Test results correlate well with several widely used batteries of tests for assessing native and nonnative speakers' proficiency in Germany, other parts of Europe, and Israel. Interestingly, they also found that C–Tests correlated very highly with the judgments of experienced classroom teachers with regard to the relative ranking of students. However, Klein–Braley and Raatz make no claims about whether the procedure can discriminate between a learner's strengths and weaknesses with regard to specific skills (Raatz 1985:60). They do not assert that C–Test results give particular insight into either the writing or reading process.

There are two recent studies that are particularly relevant to the use of C-Test for placement in Japanese college and university EFL programs. One, conducted by Akihiko Mochizuki, examined the correlation between C-Tests with three other commonly used means of assessment (1994:41-54). He used four extended C-Test passages, each employing a different rhetorical mode: narration, explanation, description and argumentation. Unlike canonical C-Test superitems, each passage was approximately 400 words long and included 120 half deletions. Mochizuki compared the results from each passage independently with the second level of the *Test of the Society of Testing English Proficiency* (STEP), the listening component of the *Comprehensive English Language Test for Learners of English* (CELT), and a dictation test of his own design.

Mochizuki found that the extended C-tests ranged from reliable to very reliable. Of these, the narration passage was highest in reliability, and it had a higher correlation with the STEP test than the others. However, he found that the correlation, using Pearson product-moment procedure, was not as high as he had anticipated. He also found that there were very low correlations between the extended C-Tests and the dictation and listening tests.

The second study, conducted by Cecilia B-Ikeguchi, was to some degree in response to Mochizuki's research (1998:3-8). Icheguchi's investigation focused on answering two questions: 1) Could both the canonical C-test and Mochizuki's extended passage C-Test discriminate between the competency levels of returnees and other students enrolled in the first year at a Japanese university?

2) Which of the two tests was superior in terms of reliability and in correlation with an external criterion? One of the tests consisted of four passages from texts that ranged in level from 6.7 to 9.6 on the Flesch–Kincaid native speaker readability index. Each passage had 25 items. The second test was based on the same narrative passage that Mochizuki had used and contained the same number of half deletions as in his study. As in Mochizuki's study, the STEP test was used as an external criterion.

The results showed that the returnee group consistently obtained higher mean scores on both kinds of C–Test which demonstrated that this procedure does discriminate between levels. The study also found that the C–Test made up of the four shorter passages had higher reliability and had a superior correlation with the STEP test than did the extended narration passage version of the C–Test. Although Ikeguchi found that the correlation was "moderate," she concludes that this, "...suggests that it is possible for C–Tests to tap different language abilities of ESL learners." (1998:7).

The Cloze /C–Test Study at Kwassui Women's College

The purpose of the study done at Kwassui Women's college was to establish whether C–Test is a more efficient way of tapping these abilities than random cloze procedure. Would C–Test prove to be a more viable way of comparing and ranking the abilities of EFL learners at a Japanese institution of higher education? To answer these questions, it seemed logical to minimize construct and procedural differences between the two tests in order to better com-

pare them.

Method

Subjects:

Data was collected from 153 students enrolled in the English Literature Department's first year English reading course (English C) between April 1998 and February 1999.

Materials:

Eight different passages were used to construct the tests. Two test formats were made for each passage: one a random reading cloze, and the other a parallel C–Test. In this way, the original content for both versions of the tests was always exactly the same. The texts were chosen from a variety of contemporary topics discussed in the international English media; however, the passages selected were specifically written and published for an ESL/EFL audience, and could stand on their own as meaningful units. These included texts that focused on climatology, destruction of the environment, education, physically challenged individuals, poverty, searching for employment, transportation, and urban planning. The passages ranged between 150 and 286 words in length. The average length of 215 words was shorter than that of most passages used for traditional random cloze testing and had fewer deletions, but this made them more similar to "superitems" in prescribed C–Tests. According to the Flesch–Kincaid native speaker readability index, the eight passages ranged from 4.1 to 9.3 in grade level, with an average of 6.4. However, the tests were not given in ascending order of reading difficulty according

to this native speaker index.

In the construction of the tests, an attempt was made to make the two versions as similar as possible. An every seventh word deletion pattern was arbitrarily chosen for the clozes. In the cloze version of each text, the first and last sentences were left intact. Then, beginning in the second sentence, approximately every seventh word was deleted. Proper nouns and numbers were not deleted. However, unlike the C-Test, one-letter words were included in the deletion pattern. A total of 20 deletions were included in each passage.

In the C-Test version of each text, the first sentence was also left intact. Then, Raatz and Klein-Braley's "rule of two" was followed, and a total of 20 words were half deleted. The one exception to the prescribed rule of two was that in four of the passages, the half-deletion pattern began on the first word in the second sentence rather than on the second word. This was done to allow for shorter appropriate cloze texts, and to permit deletion to begin at exactly the same point in both versions of the test passages. After the twentieth deletion, unaltered sentences were allowed to continue to the end of the text. Because exactly the same passages were being used for the clozes, and a cloze requires a longer reading, there were more complete sentences after the partial deletions than would normally be included in a C-test superitem. In this way, all of the test takers had exactly the same original text on which to base their hypotheses.

The results were two sets of parallel tests. Each set contained four "superitem" cloze passages with a total of 80 deletions, and four "superitem" C-test passages with 80 half-deleted words. Two superit-

ems with deletion / or half deletion beginning on the first word of the second sentence were included in each of the series of tests to minimize the differences between them.

Procedures:

Test-takers were divided into two groups, each consisting of three first-year classes. To lessen the anxiety level of the test takers, they were assured that the results had no bearing on their grades, and that the information they were providing was being collected to help their instructors know how to design better tests. In the first semester, one group (A-group) was given a practice cloze based on a text that was slightly longer and easier than those used to derive data in the study. The other group (B-group) was given a practice C-Test version of the same passage. The test administrators gave very explicit instructions to ensure that the students understood the procedures, and what they were being asked to do. Subsequent to the practice test, approximately every two weeks throughout the semester, a cloze passage was given to A-group, and a C-Test passage based on the same text was given to B-group. Each group took four tests. No explanations of the content of the materials were given in either English or Japanese. Students were asked to work through each test in 15 minutes. Although the allotted time was quite generous relative to what is normally given for C-Tests, it was felt that cloze versions may require more, and there was a desire to keep this element of the study constant.

In the second semester, the practice procedures were repeated and a series of tests based on four new passages were given.

However, this time, A-group received the C-Test version of the texts, and B-group received the cloze versions.

Scoring

The acceptable-word method of scoring was employed by the two native speakers grading the tests. It was decided that this means would more precisely reflect the range of the subjects' abilities and hence render more accurate rankings.

Results

Of the 153 students participating in this research project, 84 took all of the superitem tests for their respective groups. Only data provided by these 84 were considered in the final analysis. Statistical analysis performed on the data provided the following results:

A group	Cloze (superitems 1-4)		C-Test (superitems 5-8)	
	Mean	43.35	Mean	48.825
	Standard Deviation	9.2308	Standard Deviation	8.8256
	Range	41	Range	38
	Minimum	23	Minimum	30
	Maximum	64	Maximum	68
B group	C-Test (superitems 1-4)		Cloze (superitems 5-8)	
	Mean	54.7727	Mean	37.0454
	Standard Deviation	7.9996	Standard Deviation	9.7168
	Range	48	Range	49
	Minimum	23	Minimum	10
	Maximum	71	Maximum	59

Using the Spearman Rank Order Coefficient, it was found that there was significant correlation between cloze and C-test results (0.5372 at 39 d.f. $p < 0.001$; 0.6128 at 38 d.f. $p < 0.001$). The Spearman Rank Order Coefficient was used rather than Pearson because numbers were not normally distributed.

Discussion

The C-Tests did rank the students participating in this study similarly to the random cloze and could place students satisfactorily into groups based on relative general competency. It would also appear that C-Test could yield data in a shorter period of time than the random cloze. In this study, each superitem was administered as a separate test, but careful observation of the test-takers suggested that Raatz and Klein-Braley's allotment of seven minutes per superitem would have been sufficient time for the majority of the students participating. In contrast, it appeared that students working on the clozed versions of the superitems needed between 10 and 12 minutes despite the fact that these passages were generally somewhat shorter than those used for standard random cloze tests. Indeed, some students appeared to be still filling in blanks up to the allotted 15 minute limit. Only running both kinds of tests as multiple superitems in one sitting could actually confirm exact differences in optimal time limits.

Furthermore, based on the results of this study, it appears that fewer pilot studies would be necessary to assure reliable C-Test superitems than would be true for cloze. By definition, a reliable

norm-referenced test should center the sample around 50% correct responses and produce high standard deviation (Brown 1983:116-117). All of the passages employed in this study were being used for the first time. The C-Test superitems appear to have been reliable in that they yielded an average mean of 51.7988 with standard deviation at 8.4126. To attain two balanced C-tests centered at 50% would appear to be a simple matter of exchanging superitems between the two. However, this cannot be said of the cloze superitems. The cloze passages did produce a higher average standard deviation of 9.47, but the average mean on both tests was significantly below optimal at 41.1977. In fact, one of the more difficult superitems in the first test and two in the second test would either have to be modified or replaced if two separate reliable tests were desired (Brown 1983:116-118).

However, do such differences in average mean and standard deviation suggest that these two kinds of norm-referenced tests are not actually measuring all of the same competencies? Although the correlation between the two kinds of tests is significant, it is only moderately so. Are we therefore "comparing apples and oranges"?

There is an average difference of 11.6 when comparing the average mean score of all of the cloze passages to all of the C-Test passages. This is despite the fact that the original passages being used were exactly the same, and deletion began at the same point. Moreover, a "post-mortem" analysis of the tests shows that 22.5% of all individual items were exactly the same; 36 of the 160 words that were deleted on the clozes were the same words that were half delet-

ed, or damaged, on the C-Tests. This "collateral damage" in itself may not be that significant, but further analysis of all of the 160 items shows that the proportion of content word deletions to function word deletions was minimally different in the clozes and the C-Tests.

Although what constitutes a function word in a given context is debatable, Klein-Braley found when comparing clozes that the proportion of content word to function word deletions was one variable that could make significant differences in how difficult clozes were even when they were based on the same text (1984:98-99, 1985:43-45). It is believed that because function words belong to small closed classes, they are more easily predicted by test takers (Abraham & Chapelle 1992:469). Although no attempt had been made to control the kinds of words being selected for deletion or half deletion during construction of the tests, the overall results are remarkably similar. All cloze deletions can be broken down to 55% content words and 45% function words, while C-Test half-deletions are 58.75% content words and 41.25% function words. Therefore, in these series of tests, differences in the proportion of content to function words cannot explain why the cloze scores are consistently lower. If anything, according to the content to function word breakdown, the C-tests should be slightly more difficult. Has the construction of the C-Test neutralized this factor? If so, exactly what about their construction has done this?

Differences in the passages in the two sets of tests may help to explain the disparity in scores. How the two groups performed on the two kinds of test seem particularly apparent when one looks at the results in the second set of superitems. Group A's mean score im-

proves by 5.475 when moving from the cloze superitems in the first test to the C-Test superitems in the second test, while Group B's mean score drops by 17.7273 on the cloze version of the same test.

In contrasting the passages used in the two series of tests, two factors stand out. The average readability score using the Flesch Kincaid native speaker index of passages 1–4 was 7.3, and the average length was 179 words. However, the average readability score on passages 5–8 was considerably lower at 5.45, but the average length of the passages was higher at 250 words per passage.

Although a complete item by item analysis of all passages has not been completed at this time, these differences may confirm two findings in research related to cloze procedure. Brown (1998) has found that native speaker indices of readability are weakly related to how difficult Japanese EFL learners find passages. In addition, Bachman (1985) has found that cloze items become more difficult as context increases, and therefore the length of a passage may be a variable in the results in any given test. It seems quite possible that the length of the passages used in the second test made both kinds of tests more difficult, but cloze is perhaps more sensitive to this variable.

Even though the C-Tests appear to rank students according to their general level of competency, the competency they measure may be less "general" than that measured by cloze procedure. C-Tests are front loaded in 2 ways. First, by supplying the first half of items, C-Tests are de-emphasizing the importance of lexical knowledge and emphasizing grammatical competency. The test only requires the test

taker to complete words, and in English at least, this is often more of a grammatical task than one which requires knowledge of lexical roots. Perhaps C-Test is more a pragmatic grammar test, than a test of any other skills. Brown has found that variables such as the percentage of items over 7 letters in length, the frequency deleted items appear elsewhere in a text as words, as well as the proportion of deleted function to content words relate to how difficult a given cloze passage is for Japanese EFL learners (1998:26-30). Although these same factors may affect the difficulty of the half-deleted items in a C-Test passage to varying degrees, when Bachman's depth of context factor is added to the equation, one has to wonder whether C-Test is measuring all of the same skills that cloze does. This is pertinent to the second aspect of front loading taking place in C-Tests. Compared to clozes, C-Tests do not require the test taker to exercise productive skills beyond a fairly limited number of sentences near the beginning of a passage.

Further test by test and item by item analysis of the Kwassui data would be necessary to confirm this, but intuitively, when one compares the results on the tests constructed from passages 5-8 to passages 1-4, the difference in average length seems to be a salient factor. This would be especially true if the significantly lower native speaker readability measurements have any meaning in terms of the relative EFL difficulty levels. It seems likely that a test taker completing a cloze passage would have needed to negotiate more information, sort and track it, and exercise her short-term memory more than she would in a canonical C-Test. Therefore, a test-taker who

can complete full deletions well into the context of a cloze passage, even a passage at a relatively low level of reading difficulty, may be using skills that are not being tested when she is completing half deletions in short passages. This may also explain the differences in results between Mochizuki's longer version of C-tests and the battery of canonical superitems that Ikeguchi tested. Mochizuki's texts were longer, and required the test taker to use productive skills well into the passage.

Such skills are something that would have a bearing on academic language competency given that students are expected to read and write longer texts in their second language than C-Test passages. As students do a great deal of reading and writing during their academic careers, it seems that measuring such competencies is something that placement tests should be doing when students enter a program. Even if we were to take the importance of vocabulary out of the academic equation, being able to keep track of grammar over an extended text is important in both reading and writing. This is perhaps the difference between "theoretical" and "applied" grammar skills. Doing well on a grammar test does not mean a student can track grammar clues over a passage, or write a grammatical paper. It seems possible that canonical C-Test passages may not be as sensitive to such competencies as cloze procedure.

Conclusion

The purpose of this research was to establish which norm-referencing procedure was a more efficient and viable way of assess-

ing EFL learners in a college or university in Japan. Although C-Test proved itself capable of doing this expediently, questions have arisen regarding whether its current construction allows it to test for as many of the language competencies that apply to academic study in EFL as does the random cloze. Are both tests equally sensitive when it comes to assessing such competencies? At this point in time, a great deal more testing and analyses would be necessary to clarify this.

One way to maintain the obvious efficiency of the standardized C-Test, and at the same time better test for the competencies needed in academic work, might be to extend the rule of two to include half deletion in *every other sentence*, that is in the second, fourth, sixth, etc., sentences. This may not necessitate passages a great deal longer than those currently employed in C-Tests, as it is possible there would only be fewer sentences left intact at the end of the passage. However, it would require the test taker to exercise productive skills deeper into the context.

Furthermore, to balance the test more between grammar and lexical competencies, it might be prudent to include a series of superitems that delete the first half of words rather than the second. It seems likely that in an EFL context such a test would prove more difficult for many students than the prescribed C-Test; however, it might do more to identify and discriminate among the competencies of learners being ranked.

Such changes might satisfy the need for standardization that C-Test addresses, and at the same time test for more of the com-

petencies that cloze procedure appears to engage. More research could substantiate whether this is at all feasible, necessary, or desirable. However, it seems possible that several relatively short reliable clozes, used as superitems in a larger test, may still yield a more accurate assessment of learners skills within an academic context. In the long run, pre-testing and establishing such cloze tests may actually prove to be a more valid and efficient way of doing this. Further research would be necessary to confirm these intuitions.

Acknowledgments

This research project was *very* much a team effort. Therefore, I would like to sincerely thank the students who participated in this study, and their instructors, Richard Bent, Sergio Mazzarelli, and Andrew Gorringe whose assistance and cooperation made this study possible. Dr. Mazzarelli was very helpful with technical assistance for which I am especially grateful. In particular, I feel truly indebted to Mr. Gorringe for the considerable time, energy and input he generously gave from conception to completion of this work.

References

Abraham, R.G., & Chapelle, C. A. (1992). The Meaning of Cloze Test Scores: An Item Difficulty Perspective. In *The Modern Language Journal*, 76, 468-479.

Bachman, L. F. (1984). An Examination of Some Language Proficiency Tests from a Communicative Viewpoint. In *Practice and Problems in Language Testing*. Papers from the International Symposium on Language Testing

(8th, Tampere, Finland), 5–19.

Bachman, L. F. (1985). Performance on Cloze Tests with Fixed–Ratio and Rational Deletions. In *TESOL Quarterly*, 19, 535–557.

Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Hanscombe, R. A. Orem, & B. P. Taylor (Eds.), *On TESOL '83*, 109–119.

Brown, J. D. (1998). An EFL Readability Index. In *JALT Journal*, 20, 7–36.

Fotos, S. S. (1991). The Cloze Test as an Integrative Measure of EFL Proficiency: A Substitute for Essays on College Entrance Examinations? In *Language Learning*, 41, 313–336.

Ikeguchi, C. B. (1998). Do Different C–tests Discriminate Proficiency Levels of EL2 learners? In *Shiken: The Official Newsletter of the Japanese Association of Language Teachers Special Interest Group on Second Language Testing and Evaluation*, 2, 3–8.

Klein–Braley, C. (1984). Advance Prediction of Difficulty with C–Tests. In *Practice and Problems in Language Testing.* Papers from the International Symposium on Language Testing (7th, Colchester, England), 97–112.

Klein–Braley, C. (1985). Tests of Reduced Redundancy Theory. In *Language Testing in School.* AFinLA Yearbook. 33–48.

Lee, S. (1996). The Concurrent Validity of Cloze Test with Essay Test among Korean Students. In *Texas Papers in Foreign Language Education*, 2, 57–69.

Madsen, H. S. (1983). *Techniques in Testing*, New York, NY: Oxford University Press.

Miller, G. A. (1969). *The Psychology of Communication: Seven Essays*, Baltimore, MD: Penguin Books, Inc.

Mochizuki, A. (1994) C–Tests: Four Kinds of Texts, Their Reliability and Validity. In *JALT Journal* 16, 41–54

Oller, J.W., Jr. (1978) Pragmatics and Language Testing. In B. Spolsky (Ed.) *Approaches in Language Testing*, CAL, 39– 57.

Raatz, U. (1984). The Factorial Validity o f C–Tests. In *Practice and Problems in Language Testing.* Papers from the International Symposium on Language Testing (7th, Colchester, England), 124 –139.

Raatz, U. (1985). Tests of Reduced Redundancy The C–Test, A Practical Example. In *Language Testing in School.* AFinLA Yearbook, 49–62.

Raatz, U. & Klein-Braley, C. (1981). The C-Test A Modification of Cloze Procedure. *Practice and Problems in Language Testing.* Papers from the International Symposium on Language Testing (4th, Colchester, England), 26-52.

Silberstein, S. (1989). *Techniques in Teaching Reading*, Unpublished manuscript. Chapter 9, 32-44.

Spolsky, B. (1969). *Reduced Redundancy as a Language Testing Tool.* Paper read to the Language Testing Section of the 2nd International Congress of Applied Linguistics, Cambridge, England

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 414-438

Taylor, W. L. (1957). Cloze readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology*, 41, 19-26

Valette, R. M. (1977). *Modern Language Testing*, New York, NY: Harcourt Brace Jovanovich, Publishers, Inc.

Received January 31, 2000