Bulletin of Fukuoka University of Education, Vol.57, Part Ⅰ, 47—57 (2008)

Can Cloze Tests Measure Discourse Competence in ESL/EFL Appropriately?

高 梨 芳 郎

Yoshiro TAKANASHI
英語教育講座

（平成19年10月１日受理）

Although cloze tests have been used as economical measures of EFL/ESL proficiency, one of the central questions concerning test construction has not been answered yet: Can cloze tests measure discourse competence in ESL/EFL appropriately? This article deals with this issue and proposes an alternative discourse cloze test. First, we examine the following three questions: (1) Are cloze items sensitive to constraints across sentences? (2) Are all cloze tests equivalent in the same text? (3) Can the C-Test take the place of the standard cloze test? Second, we review the construct validity of the rational cloze test and previous discourse cloze tests. Third, we propose both a more authentic, discourse-conscious cloze test and a conversational cloze test for EFL/ESL contexts.

INTRODUCTION

The cloze procedure has been used as a quick, efficient measure of language proficiency since Taylor (1953) developed it to measure the readability of a passage of prose. A cloze test is constructed by deleting words from the text and requires the student to fill in the blanks. The test requires the student to understand the whole text, predict the missing words, and restore the text. The process is not unlike what native speakers do in receiving and sending messages in real communication. Thus, the cloze procedure has been shown to be an economical, valid and reliable measure of overall proficiency in ESL/EFL (Brown, 1980, 1983; Fotos, 1991; Oller, 1971, 1972, 1973; Oller & Conrad, 1971; Stubbs & Tucker, 1974). The early validity studies of cloze tests indicated that the range of correlation coefficients reported between cloze tests and criterion measures was from .63 to .89 (Brown, 1980). The reliability of the cloze test has been shown by the estimated coefficients which ranged from .61 to .95 for various scoring methods
(Brown, 1980).

Various types of cloze tests have been developed. The most traditional one is the standard cloze test, where the *n*th word is deleted (called fixed-ratio method) and the student is required to fill in gaps to restore the text (Oller, 1979). A second technique is the multiple-choice type, where the examinee is provided with alternative answers from which to select the correct answer for each gap (Jonz, 1976). This method reduces the difficulty in marking, while the intended test validity confines itself to receptive skills (Brown, 1980; Porter, 1976). The variable-ratio method has another possibility to delete just the words that are richly laden with meaning with discretionary judgment (Oller, 1979). Bachman (1985) proposed a rational cloze test ('a gap-filling test': Alderson 2000; Yamashita 2003) in which four types of cloze items were prepared according to the amount of context needed to fill in the blank. The last one is the C-test in which the second half of every second word is deleted in a number of short texts (usually five or six) (Klein-Braley & Raatz, 1984). It is suggested that the C-test is the most valid and reliable test in that it produces a random sample of the text (Klein-Braley, 1983, 1985; Klein-Braley & Raatz, 1984).

Although early research into cloze tests promised that the cloze method could be an economical, reliable, and valid measure of ESL/EFL proficiency, the standard cloze test has been seriously attacked concerning its construct validity. Among the questions are: What construct can the cloze test measure? Different deletion rates and starting points applied to the same text can produce cloze tests which differ considerably in their difficulty, reliability, and validity (Alderson, 1983; Klein-Braley, 1983; Klein-Braley & Raatz, 1984; Lado, 1986; Porter, 1978). This question relates closely to another question: Are cloze items sensitive to discourse constraints across sentences? Cloze tests can measure only lower order skills if the results of the tests markedly differ according to the passage difficulty and deletion ratio (Alderson, 1983; Porter, 1978). This study attempts to answer the questions above and argues for the use of discourse cloze tests to tap discourse competence in ESL/EFL.

WHAT CONSTRUCT CAN THE CLOZE TEST MEASURE?

Are Cloze Items Sensitive to Constraints across Sentences?

There have been two ways to investigate the effects of context on cloze scores. One was to compare the performance of students on the standard cloze version and on a scrambled version of the same text. The other was to investigate the effect of varying amounts of context on the predictability of cloze items.

Chihara et al. (1977) administered a standard cloze test and a scrambled version of the same text to 201 Japanese students of English and 41 native English speakers, and found that the scores on the scrambled version were significantly lower than the scores on the standard cloze test. They concluded that the standard cloze test is sensitive to constraints across sentences. Chavez-Oller et al. (1985) confirmed Chihara et al.'s (1977) findings by using item analysis, and showed that almost all the items in the scrambled version were significantly lower than those of the original version. Ten per cent of the cloze items were found to be sensitive to long-range constraints across sentences.

Porter (1983) gave 150 subjects 8 sets of 12 sentences with a blank plus a bilateral context of 5-12 words, and determined that a context quantity had not much effect on acceptable predictability of a blank word. He suggested that since quantity of context beyond 5 or 6 words bilaterally did not show much contribution to the predictability of a deleted word, cloze tests may not sensitive to wide-ranging context beyond 5 or 6 words.

Thus these research studies reached contradictory conclusions. However, they seem to suffer from difficulties in their experimental procedures. The former studies (Chavez-Oller et al. 1985; Chihara et al. 1977) would have had a different result if the acceptable words had been counted as correct. In the scrambled version, filling in the gap with the exact word (especially in the case of a content word) is more difficult because it has less contextual information, while filling in the gaps with acceptable words is not so difficult because there is a much wider choice of acceptable words than in the original version:

He didn't HAVE a suit made, though, because his (   ) wouldn't let him order one.
Exact word: FATHER; Acceptable words: FATHER, MOTHER, PARENTS ....
(Chavez-Oller et al. 1985, p. 203)

There are several questions which need to be asked of Porter's (1983) experiments: First, the segments used seem to be unnatural in that the mean scores of the native speakers were 48.5% (exact word scoring) and 66.6% (acceptable word scoring), where one would have supposed that the scores would approach 100% among native speakers. Second, the 12 segments which made up each set of a bilateral context are, in fact, a relatively small number as representatives of each set of a bilateral context. Lastly, each blank word was not controlled in terms of word frequency which might have significant effects on cloze difficulty.

Are All Cloze Tests Equivalent in the Same Text?

Three major investigations (Alderson, 1983; Klein-Braley, 1983; Porter, 1978) have shown that different starting points and deletion rates applied to the same text can result in cloze tests which differ considerably in their difficulty, reliability, and validity.

Alderson (1983) experimented with the effect of changes in deletion rates (6th, 8th, 10th, and 12th) and text difficulty (easy, medium, and difficult) on the test validity. Twelve cloze tests were constructed from three texts, 650 words in length, with 50 words deleted. Each cloze test was administered to 30 different nonnative learners of English and was validated with the ELBA test. The results showed that the validity coefficients could vary considerably with the changes in deletion rates and text difficulty. Thus, he concluded that 'the cloze would seem to be very sensitive to the deletion of individual words,' and that 'one must ask whether the cloze is capable of measuring higher-order-skills' (Alderson, 1983, p. 211).

Porter (1978) examined the equivalence of two cloze tests which were identical except that in one cloze test deletion occurred one word earlier than in the other. Each cloze test had two subtests, literary and non-literary subtests. The deletion rates were 8 words and 50 blanks were prepared in each subtest. The first and second tests were administered to the same 39 university students in Poland after a three-week interval. The results showed that the two cloze tests were not equivalent, since the correlation between them was 'relatively low' (.57) and the difference in the mean scores between them was significant (F = 30.35, p<.01, means: 50.89 and 56.33, SDs: 5.72 and 6.96, respectively), all in exact scoring.

Klein-Braley (1983) also cast doubt on the question of cloze equivalence across tests. Six groups of German university students majoring in English were asked to complete two different cloze tests. The scores were correlated between each group of students. The correlations ranged from .39 to .70, and the reliability coefficient (K-R 20) of each cloze test was .15 to .74, all in exact scoring. Thus, Klein-Braley (1983) insisted that 'a cloze is *not* a cloze', and that 'there is no such thing as cloze equivalence across the tests' (p. 226).

The assumption that 'any two individual cloze tests are equivalent tests' (Klein-Braley, 1983, p. 220) seems naive, since there is the least possibility that any passage selected from any text produces a random sample of all possible elements of the language. The question of cloze equivalence across tests should be examined within the particular text selected for the test. Furthermore, the experimental procedures employed by critics of cloze tests should be questioned.

The cloze tests employed by Alderson (1983) might have tested different elements of the language to the subjects, since the texts consisted of passages of different lengths: 300 to 600 word fiction texts (language items tend to fluctuate easily in fiction). The significance of differences in the correlations also needs to be tested. Although the experiment conducted by Porter (1983) appears to be more rigorous, two questions need to be asked about his experimental procedure: (1) there seems to have been a strong practice-effect on test 2, especially on non-literary subtest B, since the blank words in test 2 would have been easy to remember in that they appeared just one word before the blanks in the test 1, which had been tested 3 weeks before, and that the content of the subtest B was more familiar and memorable. This shows there was some possibility that the gap filling abilities needed in test 1 and test 2 differed significantly. (2) The number of subjects was rather small (39) and thus the range of the tests (especially on subtest A: test 1: 22-31, test 2: 21-34) was narrow, which could make the correlation between the two tests relatively low. This problem also applies to Klein-Blaley's (1983) experiments. The number of subjects was 23-53, 8 out of 12 under 32, and the number of blanks in each test was 30-50, half of them under 40. These cloze tests were not necessarily suitable to examine the question of cloze equivalence. Thus, the question of cloze equivalence still awaits a more definitive answer.

Can the C-Test Take the Place of the Standard Cloze Test?

The C-Test, developed by Raatz and Klein-Braley in 1981, is a reduced redundancy test which deletes half of every second word in a text (the rule of 2). A C-Test consists of several different texts (usually five or six texts), each of which usually has around 25 deletions (Klein-Braley, 1997; Klein-Braley & Raatz, 1984). The C-Test claims to eliminate the problem that different deletion rates, starting points, and scoring methods can affect cloze scores, since only two C-Tests can be constructed in the same text and only exact scoring is allowed. Klein-Braley and Raatz (1984) summarized the research into C-Tests and reported that most of the investigations showed high reliability (more than .80) and validity (more than .50) for different groups and languages. This trend has been supported by other researchers (Babaii & Ansary, 2001; Dörnyei & Katona, 1992; Grotjahn, 1986; Negishi, 1987).

However, there are still many other researchers who question the use of C-testing (Carroll, 1986; Cohen, Segal, and Weiss, 1984; Jafarpur, 1995; McBeath, 1989, 1990; Piper 1983). Carroll (1986) cautioned that the C-Test 'seems to be limited to the measurement of general proficiency, chiefly at lower level of ability, in written language' (p. 128). McBeath (1989) concluded that 'while C-Testing may be a legitimate device of L1 testing, it lacks a theoretical basis for application with FL learners' (p. 36). Bradshaw (1990) showed that the C-Test was rated most negatively by subjects with different language proficiency. Jafarpur (1995) constructed 20 C-Tests with different ratio and/or deletion start, and pointed out that 'various deletion ratios and deletion starts produce different tests' (p. 209). Piper (1983) administered the C-Test and cloze test to the same subjects, and found that the cloze test correlated better with the grading test (.94) than the C-Test (.79).

Is the C-Test superior to the cloze test? Empirical evidence in support of each test is scanty and difficult to obtain, as shown in Jafarpur's (1995) investigation of 20 C-Tests with different ratio and/or deletion start. The number of items (60% of items are less than 20) and subjects (all 15-19) in each C-Test administered to non-native speakers was too small to reach some desired reliability and validity coefficients. The result of Piper's (1983) experiment is also inconclusive, since the present author found that there was no significant difference of the correlations (t = .80, p = .42). Moreover, the cloze test and the C-Test with the same words deleted, constructed from the same text showed no significant difference in terms of validity and reliability (Takanashi, 1995). These tests, thus, should be examined to see whether they can measure communicative competence appropriately, especially the ability to process discourse.

APPROACHES TO MEASURE DISCOURSE COMPETENCE

Can the Rational Cloze Test Measure Discourse Competence Appropriately?

In order to examine whether cloze tests are capable of measuring syntactic and discourse levels of competence, Bachman (1982) developed the rational cloze test in which deletions were made not based on systematic *n*th deletion but based on the level of language context: '1) syntactic, which depended only on clause-level context, 2) cohesive, which depended upon the interclausal cohesive context, and 3) strategic, which depended on parallel patterns of coherence' (p. 63). He used confirmatory factor analysis and suggested that the cloze test with rational deletions can be used to measure textual relationships beyond clause boundaries.

Bachman (1985) revised these categories regarding the level of language context to make cloze items easier, and proposed four types of deletions: 1) within clause; 2) across clause, within sentence; 3) across sentences, within text; and 4) extra-textual. He constructed the rational cloze test with this rational deletion procedure, and compared it with a standard cloze test constructed from the same text. Results showed that 60% of the total items of the rational cloze test formed types 2 and 3, while only 10% of the total items of the standard cloze test formed types 2 and 3. Compared with the standard cloze test, the rational cloze test tended to decrease the number of correct answers for each type of items in proportion to the amount of context: Type 1 > Type 2 > Type 3 > Type 4. These results suggested that cloze procedure with rational deletions can be used to measure higher order skills (coherence and cohesion).

What construct the cloze test can measure seems to depend on the quality of cloze items. If we are to measure higher order skills with fewer items, we should focus on syntactic and textual functions in a given text. Systematic *n*th deletions can ignore the syntactic and semantic relationships in a text, and are therefore likely to produce inconsistent results. Rational deletion procedures may be promising for constructing cloze items sensitive across sentences. However, the deletion criterion Bachman (1985) proposed needs to be improved in terms of discourse constraints if we intend to measure discourse competence appropriately. It will be necessary to examine the syntactic and semantic structure of cloze passages by using discourse analysis. Discourse cloze tests should include cloze items sensitive to linguistic and cohesive structures of the target language.

Approaches to Discourse Cloze Tests

In addition to knowledge of structure and vocabulary, we need knowledge of how the sentences relate to each other so as to understand the meaning of texts. This knowledge relates to text-forming devices whose description and analysis are found in Halliday and Hasan (1985, p. 82): 'grammatical cohesive devices (reference, substitution and ellipsis, and conjunction) and lexical cohesive devices (general, instantial, and continuatives)'. Among these categories, Halliday and Hasan (1976) showed that two major categories of lexical cohesion are reiteration and collocation. Hoey (1983) also insisted that 'the majority of the sentences in discourse are connected unambiguously with their neighbours by anaphoric devices of several kinds (e.g., *such*, *its*, *this*) and by simple repetition' (p. 6). Hoey (1991) argued that lexical cohesion is the single most important form of cohesion, accounting for 42 - 48% of cohesive ties in texts (p. 9). The background knowledge of the reader or listener plays a more obvious role in the perception of lexical relationships than in the perception of other types of cohesion. Collocational patterns, for example, 'will only be perceived by someone who knows something about the subject at hand' (Nunan, 1993, p. 30). Coherence defined as 'the relationship between illocutionary acts' (Widdowson, 1978, p. 28) is also a key concept in understanding spoken and written discourse. McCarthy (1991) described discourse analysis as the study of spoken and written interaction. The study of discourse analysis will contribute to 'a better understanding of exactly how natural spoken and written discourse looks and sounds' (McCarthy, 1991, p. 12). The ideas and key concepts in discourse analysis will contribute to selecting the criteria for item deletion in constructing discourse cloze and interpreting the results of tests.

Towards an authentic discourse cloze, Deyes (1984) proposed that communicative units should be considered more appropriate deletion items than single words, and that the selection of such items could be done on the basis of recoverability and relevance of such units. He suggested that deletion items include such communicative units as thematic, transitional, and rhematic items (see Appendix A). He administered 'discourse cloze' to students and found that whereas thematic deletions and transitional items were acceptably recovered with high percentages (75% and 95%, respectively), rhemes were found to be replaced with a variety of degrees of success (mean percentage of correct: 39%) (p. 130). Recoverable rhemes were those whose content could be derived by knowledge of collocations and/or by common knowledge of the world. He also commented that relevance could be examined by judging to which degree the communicative units place demands on students' world knowledge, and whether 'their place in sentences is relevant to, or of major importance to, the discourse as a whole' (p. 133).

Meyer and Tetrault (1987) proposed the use of cloze-like activities for reading exercises, in which words were deleted 'according to their situational and contextual motivation' (p. 410). In order to help students develop foreign language reading strategies, they suggested 'deleted items must be chosen primarily for their cohesive or situational connections outside the immediate sentences in which they appear' (p. 414) (see Appendix B). Their cloze-like activities also include predicting the contents of the missing parts of a text based on what students have read or what they already know, and completing blanks that contain more than one word by using their knowledge of topic-comment and discourse constraints through the text and by referring to the choices.

These discourse-conscious cloze tests seem to be promising; (1) They are designed to reflect the ability to process the text and use the contextual and co-textual clues. (2) Deletion items are not necessarily limited to those drawn from the cohesive systems, since they can also be focused on discourse as a whole. (3) Words are deleted not on a random basis (deleting every nth word) but according to their communicative values to the discourse as a whole and their recoverability in terms of being relevant to students' world knowledge as well as knowledge of collocations.

However, in order to construct a more appropriate discourse cloze, it can be advocated that some points should be improved. First, we need more empirical evidence for the suitability of such a test, even if the theory underlying 'discourse cloze' seems to be legitimate. We should determine how far 'discourse cloze' is suitable to measure discourse competence. Unfortunately, such evidence involved in concurrent validity would take time to obtain, since a standardized measure of discourse competence has not yet been found in the field of EFL/ESL testing.

Furthermore, recoverable communicative units, especially rhemes, are supposed to be those whose contents can be predicted by knowledge of collocations and/or by students' world knowledge. However, such cloze items can make the test heavily reliant on grammar and/or schema (frame) which is less related to the ability to follow discourse structures (e.g., ' ... People usually have large and beautiful gardens so that they can spend *their leisure time* outside') (Deyes, 1984, p. 133). These deletion words (underlined in the example) will be easily inferred appropriately from the phrase, 'spend ... outside'). If we are to measure the ability to process the co-textual clues as well as the contextual ones, such discourse cloze should include not only word deletion of varied contexts, but *sentence* deletion of varied co-text, considering the recoverability and the shared background knowledge in terms of discourse structures. Scoring would be another problem in discourse cloze testing, since word deletion is applied to longer units than in classical cloze. Having longer units also increases the range of word choices and combinations of word choices; this in turn potentially renders scoring more difficult. It also increases the possibility of other errors, e.g., in syntax, if testees are writing whole clauses or sentences rather than single words. Criteria regarding such issues as acceptability of word combinations, clause elements, clause relations and coherence relations would need to be made explicit. While discourse cloze provided with choices can be scored easily, the appropriate scoring method for tests without choices usually requires that the marker knows all the acceptable answers for each blank. However, regarding acceptable word scoring it has been demonstrated that although it is more complex than the exact word method, it has the advantage of being a more valid and fairer method of scoring (Bachman, 1982, 1985; Brown, 1980; Oller, 1972, 1979). Thus, this method could be recommended for scoring discourse cloze. This can be accomplished with the help of native speakers of the target language.

The final suggestion is related to the use of 'discourse cloze', which was originally intended to reflect the *reader's* ability to follow information through discourse clues: Such cloze without choices could be modified to create a more suitable discourse cloze test. This could also tap the interactive abilities to understand and express the writer's idea to use the textual and discourse features of target language if we can prepare cloze items (including both deletion words and deletion sentences) appropriately. If we are to measure just higher-order-skills across sentences, it will be necessary to open our 'clozed minds' (Meyer and Tetrault, 1986), and deletion items should include not only words of varied context but *sentences* of co-texts, depending on which abilities are to be measured. Moreover, considering the fact that text selection for 'discourse cloze' does not involve discourse from authentic conversations, another possibility for modifying 'discourse cloze' would be authentic conversational cloze tests whose communicative units (sentences) are deleted to measure the conversation skill by using transcripts of authentic conversation. This will make the conversational cloze test which Brown (1983) proposed a more authentic and discourse-conscious one. It will be useful for evaluating students' oral communication skills in writing.

Examples of Discourse Cloze Tests

Examples of discourse cloze are given below. Discourse cloze is usually based on an authentic text which has one or more paragraphs. The beginning of the text is left intact in order to enable testees to get the overall idea of the text. Words (sentences) are deleted according to communicative values and recoverability of the content, making reference to shared common knowledge. The words and/or sentences underlined are deletion items. Example (1) is a sample of discourse cloze tests. Example (2) is an illustration of conversational (discourse) cloze. Both of these need an acceptable scoring system. A discourse cloze whose deletion items are selected from cohesive devices is also shown in Appendix C.

Example (1) Discourse Cloze Test

The nuclear family is generally a conjugal unit. By this is meant that, in most societies, the family grows out of the union of a man and a woman who have entered into a marriage partnership. It consists primarily of a father, mother and offspring. (1) The consanguine family, by way of contrast, comprises a nucleus of blood relations plus associated spouses. (2) When someone marries, he or she is incorporated into the parental families and potentially is able to share a common life with all blood relatives. The former family type is a concentrated culture pattern; the latter is a diffused one.(3)

Because the conjugal family is a self-contained unit and dependent for progress on its resources, relationships among its members can be very intimate.(4) This allows greater freedom of expression for individual personality, but when relationships break down the consequences for individual members, especially children, can be traumatic. Disruption of the consanguine family, contrarily, is virtually impossible. Members usually consider collective responsibility more important than individuality(5), and as a result the extended family is usually stable, conservative and traditional.

Example (2) Conversational Cloze Test

A: Good morning Mr Plant. Do sit down.
B: Thank you.(1)
A: First of all I'd like you to tell me a bit about what you've been doing.
B: Well, I left school after I'd done my A levels.(2)
A: What subjects did you take?
B: French, German and Art.(3)
A: Art?

B: Well, I really wanted to study art.(4) But a friend of my father's offered me a job. He's an accountant in the City.

A: I see. In your application, you say that you only spent nine months with this firm of accountants. Why was that?

B: Well to be quite honest, I didn't like it(5) — so I got a place at the Art College.

A: Did your father mind?

B: Well, he was quite disappointed at first.(6) He's an accountant too, you see.

A: Have you any brothers or sisters?

(Source: Abbs, Cook, & Underwood, 1979, pp. 49-50.)

CONCLUSION

While cloze tests have gained in popularity among language teachers and researchers because of their simplicity in construction and administration of the tests, their construct validity soon began to be seriously questioned. What constructs can a cloze test measure? Are cloze items sensitive to discourse constraints across sentences? This paper attempted to answer these questions. It also emphasized the use of discourse cloze tests to tap discourse competence in ESL/EFL.

The most promising alternative to classical cloze is found to be selective deletion of cloze items. This offers the possibility of greater control over test items in terms of the level of item difficulty. Selective deletion also offers the possibility of focusing on items of particular areas of language and functions. Thus, discourse cloze could be used to ensure that deletion items tap the processing abilities of the macro-structure as well as the micro-structure of the target language. Another possibility would be to use various types of 'cloze-like exercises' (Meyer and Tetrault, 1987). These would have the possibility of being effective integrative tasks that allow students to deal meaningfully with authentic language.

However, some questions remain unanswered: Are discourse cloze tests valid and reliable measures of discourse competence? Do they foster students' discourse competence? These questions still await answers through strict empirical studies.

REFERENCES

Abbs, B., Cook,V., & Underwood, M. (1979). *Realistic English dialogues.* Oxford: Oxford University Press.

Alderson, J. C. (1983). The cloze procedure and proficiency in English as a foreign language. In: J. Oller, Jr., (Ed.), *Issues in language testing research* (pp. 205-217). Massachusetts: Newbury House.

Alderson, J. C. (2000). *Assessing reading.* Cambridge:Cambridge University Press.

Babaii, E., & Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle? *System,* 29,2, 209-219.

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

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly,* 19,3, 535-556.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing*, 7,1, 13-30.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal*, 64,3, 311-317.

Brown, D. (1983). Conversational cloze tests and conversational ability. *English Language Teaching Journal*, 37,2, 158-161.

Carroll, J. B. (1986). LT+25, and beyond? Comments. *Language Testing*, 3,2, 123-129.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J.W. Jr. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning*, 35,2, 181-206.

Chihara, T., Oller. J. W. Jr., Weaver, K.A., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning*, 27,1, 63-73.

Cohen, A. D., Segal, M., & Weiss, R. (1984). The C-test in Hebrew. *Language Testing*, 1, 221-225.

Deyes, T. (1984). Towards an authentic 'discourse cloze'. *Applied Linguistics*, 5,2, 128-137.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing*, 9,2, 187-206.

Fotos, S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examination? *Language Learning*, 41,3, 313-336.

Grotjahn, R. (1986). Test validation and cognitive psychology: some methodological considerations. *Language Testing*, 3, 159-185.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. New York: Longman.

Halliday, M. A. K., & Hasan, R. (1985). *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Hoey, M. (1983). *On the surface of discourse*. London: Allen and Unwin.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Jafarpur, A. (1995). Is C-testing superior to cloze? *Language Testing*, 12,2, 194-216.

Jonz, J. (1976). Improving on the basic egg: The M-C cloze. Language Learning, 26,2, 255-265.

Klein-Braley, C. (1983). A cloze is a cloze is a question. In: J. Oller, Jr., (Ed.), Issues *in Language Testing Research* (pp. 218-228). Massachusetts: Newbury House.

Klein-Braley, C. (1985). A cloze-up on the C-test: a study in the construct validation of authentic tests. *Language Testing*, 2, 76-104.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, 14,1, 47-84.

Klein-Braley, C. & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing*, 1,2, 134-146.

Lado, R. (1986). Analysis of native speaker performance on a cloze test. *Language Testing*, 3,2, 130-146.

McBeath, N. (1989). C-tests in English: Pushed beyond the original concept? *RELC Journal*, 20,1, 36-41.

McBeath, N. (1990). C-tests — some words of caution. *English Teaching Forum*, 28,2, 45-46.

McCarthy, M. (1991). *Discourse analysis for language teachers*. Cambridge: Cambridge University Press.

Meyer, R. & Tetrault, E. (1986). Open your clozed minds: using cloze exercises to teach foreign language reading. *Foreign Language Annals*, 19,5, 409-415.

Negishi, M. (1987). The C-test: an integrative measure? *The IRLT Bulletin*, 1, 3-26.

Nunan, D. (1993). *Introducing discourse analysis*. Penguin English.

Oller, J. W. Jr. (1971). Dictation as a device for testing foreign language proficiency. *English Language Teaching*, 25,3, 254-259.

Oller, J. W. Jr. (1972). Scoring methods and difficulty levels for cloze tests of ESL proficiency. *Modern Language Journal*, 56,3, 151-158.

Oller, J. W. Jr. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning*, 23,1, 105-118.

Oller, J. W. Jr. (1979). *Language tests at school*. New York: Longman.

Oller, J. W. Jr., (Ed.) (1983). *Issues in language testing research*. Massachusetts: Newbury House.

Oller, J. W. Jr., & Conrad, C. A. (1971). The cloze techniques and ESL proficiency. *Language Learning*, 21,2, 183-195.

Piper, A. (1983). A comparison of the cloze and C-test as placement test items. *The British Journal of Language Teaching*, 21,1, 45-51.

Porter, D. (1976). Modified cloze procedure: a more valid reading comprehension test. *English Language Teaching Journal*, 30,2, 151-155.

Porter, D. (1978). Cloze procedure and equivalence. *Language Learning*, 28,2, 333-341.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In: A. Hughes & D. Porter. (Eds.) (1983). *Current developments in language testing* (pp. 63-74). New York: Academic Press.

Stubbs, J. B. & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal*, 58, 5-6, 239-241.

Takanashi, Y. (1995). A comparative study of the standard cloze test, the matching cloze test, and the C-Test. *Bulletin of Fukuoka University of Education*. 44,1, 41-54.

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

Widowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Yamashita, J.(2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers, *Language Testing*, 20,3, 267-293.

APPENDIX A Discourse Cloze

Brisbane, which is the capital of the Australian state of Queensland, has a more relaxed atmosphere than Sydney, perhaps because of its pleasant sub-tropical climate. Its situation is not as impressive as Sydney's, but (1) the broad Brisbane which runs through the city centre, is full of ocean-going boats, ferries — and small boats as well. The way of life is probably the most pleasant and relaxed that you will find anywhere in a big city. People usually have large and beautiful gardens so that they can spend (2) their leisure time outside ....

(Deyes, 1984, p. 133)

APPENDIX B Cloze-like Activities

South America. It's not foreign to us.

After 13,000 flights and one million passengers, we're certainly not new to South America. In fact, we fly(1) to more cities in South America from more(2) cities in North America than any other airline. With Eastern(3), one phone call can book you to 12 cities(4) in South and Central America(5), making Lima as easy as(6) Portland ....

(Meyer & Tetrault, 1986, p. 410)

APPENDIX C Discourse (Cohesion) Cloze Test

The nuclear family is generally a conjugal unit. By (1) is meant that, in most societies, (2) family grows out of the (3) of a man and a woman who have entered into a (4) partnership. (5) consists primarily of father, mother and offspring. The consanguine (6), by way of contrast, comprises a nucleus of (7) relations plus associated spouses. When someone (8), he or she is incorporated into the parental (9) and potentially is able to share a common life with all blood (10). The (11) family type is a (12) culture pattern; the latter is a diffused (13). (14) the conjugal family is a self-contained (15) and dependent for progress on (16) resources, relationships among its members can be very intimate. (17) allows greater freedom of expression for (18) personality, (19) when relationships break down the consequences for individual members, especially (20), can be traumatic. Disruption of the (21) family, (22), is virtually impossible. Members usually consider collective responsibility more important than (23), (24) as a result the extended family is usually (25), conservative and traditional.

Answer Keys

1. this (Reference)
2. the (Reference)
3. union (General)
4. marriage (General)
5. It (Reference)
6. family (General)
7. blood (General)
8. marries (General)
9. families (General)
10. relatives (General)
11. former (Reference)
12. concentrated (General)
13. one (Substitution & Ellipsis)
14. Because (Conjunctives)
15. unit (General)
16. its (Reference)
17. This (Reference)
18. individual (General)
19. but (Conjunctives)
20. children (General)
21. consanguine (General)
22. contrarily (General)
23. individuality (General)
24. and (Conjunctives)
25. stable (General)