FACTORS AFFECTING CLOZE TEST PERFORMANCE

Daniel Jenks

1. INTRODUCTION

The cloze test has been described as a measure of overall language proficiency, whose generally high reliability and validity can be affected by numerous factors. This article explores the influence of several variables on the cloze performance of two groups of low-level English students. A different test was given to each group, although both were based on the same text, and each was marked using two different scoring methods. Performance on the cloze test was also compared to performance in a previously-completed class test.

The resulting data showed that the scoring methods used had little effect on the students' relative performance, but that the two different cloze tests correlated differently with the class test. In addition, performance levels varied according to the types of items included in the cloze tests. Reliability estimates for the two tests varied depending on the method of calculation, but scoring method had little effect.

2. PREVIOUS CLOZE TEST RESEARCH

2.1 The cloze test

Although there are many different forms of cloze tests, all are based on the concept of removing words from a text, and asking respondents to restore these missing items, using their comprehension of what remains of the text and their knowledge about the language to guide them. Words may be removed by a system

of rational deletion (deleting all examples of a particular part of speech, or verb form, for example), or by fixed ratio deletion (deleting every nth word regardless of what it is) (Richards and Schmidt, 2002:78). Different methods may also be used to score cloze tests, including accepting only the exact word deleted from the text (the EXACT method), and accepting any answer that is semantically acceptable in the context of the passage (the SEMAC method) (Owen, et al., 1997:41). Cloze tests can be used for a number of purposes, from testing readability of a text and reading comprehension of native language speakers, to testing language proficiency in second language speakers (Alderson, 1979:220). Although cloze tests are generally drawn from written texts, transcribed speech has also been used in an attempt to measure oral ability (Hughes, 2003:191).

2.2 What do cloze tests measure?

Although "the precise abilities measured by a given cloze test remain in question" (Abraham and Chapelle, 1992:468), the approach has been credited as a method of easily producing practical integrative tests, and with providing a measure of general language proficiency (Owen, et al., 2002:39) by testing a range of skills that includes "knowledge of vocabulary, grammatical structure, discourse structure, reading skills and strategies, and an internalised 'expectancy' grammar" (Brown, 2001:393). It is believed that these skills are the same as those that language users would utilise in any language context (Abraham and Chapelle, 1992:468). In essence, cloze tests for second language learners highlight the correspondence between their knowledge of language rules and the rules as they actually exist (Aitken, 1977:65).

Some have claimed that restoration of cloze test items is dependent mainly on syntactical constraints, and that only information local to an item is considered when attempting to restore it (Alderson, 1992:225). However, other research indicates that

more long-range constraints (including lexical selection and textual cohesion) may affect responses (Jonz, 1990:72). In addition, performance may be influenced by "contextual features of the test method, such as text difficulty relative to the group tested, text topic, deletion ratio, and method of student response." (Abraham and Chapelle, 1992:469).

The nature of words deleted from a cloze test may influence respondents' ability to complete it. Function words representing little lexical information are generally easier to restore than content words representing more information (Aitken, 1977:64). "Closed" items are more easily predicted than "open" items (Jonz, 1990:73). When testing non-native speakers, deleted items may not be part of the test-takers vocabulary, making them difficult to restore accurately (Kobayashi, 2002:572). However, words that appear frequently in a test are easier to restore, even if the test-taker is unfamiliar with them (ibid.).

2.3 Methods of constructing cloze tests

There a numerous variables in the construction of a cloze test that may affect its difficulty and even the areas of language proficiency that it measures (Abraham and Chapelle, 1992:474), and opinions vary on what is the most effective method of cloze design. Using the fixed ratio method, "cloze tests made by deleting every fifth word measure skills closely-related or identical to those measured by conventional multiple-choice reading or comprehension tests" (Bormuth, 1969:365). However it is claimed that rational deletion can produce tests that are equally reliable and valid, but which also enable tests to be shorter and easier to score (Bachman, 1985:550). The latter method may be preferable in many cases, because "deletion of every nth word almost always produces problematic items (for example, impossible to predict the missing words)" (Hughes, 2003:190). However, the non-selective nature of the

fixed-ratio method is one of its main advantages, as it results in a more representative range of items (Oller, 1972:151).

Even with a chosen method of deletion, entirely different tests may be created. Fixed ratio deletion using "different starting points for deletion and different deletion rates might contribute to significant variations in the nature of cloze items produced from any one text" (Kobayashi, 2002:581). As "the deletion of function words detracts less from comprehensibility than the deletion of content words" (Oller and Inal, 1971:315), if "a high number of content words happen to be chosen by fixed-ratio, results will be different to rationally deleted function words" (Abraham and Chapelle, 1992:474).

2.4 Methods of marking cloze tests

Many studies have shown that EXACT and SEMAC scoring methods produce results which correlate very highly for a given cloze test (Stubbs and Tucker, 1974:241). Therefore, due to the level of subjectivity introduced and the extra effort required by the SEMAC method, some have claimed that "the most valid and economical results are obtained by scoring correct only those responses exactly matching the deleted words" (Bormuth, 1969:360).

This may be true for native speakers taking cloze tests, however some research has shown that the SEMAC method is not only more reliable when testing second language learners, it also seems more "intuitively correct" (Brown, 1980:311) and fairer despite the potential for marker unreliability (Aitken, 1977:61). Insisting on exactly-correct answers may make cloze tests too difficult for many L2 speakers (Oller, 1972:151). Indeed, the cloze scores of low-proficiency learners show greater increases when switching from EXACT to SEMAC scoring than the scores of high-proficiency learners (Kobayashi, 2002:576). Finally, due to its inclusion of cohesive

constraints, SEMAC scoring better distinguishes between low and high level speakers (Bachman, 1982:66), making this method preferable in tests designed to highlight differences in ability (for example, school tests used as a basis for grading students).

2.5 Reliability and validity of cloze tests

Although there is a greater chance of scorer unreliability when using SEMAC scoring (Oller, 1972:151), it is generally believed that SEMAC is the more reliable of the two main methods used (Owen, et al., 1997:42). It has been suggested, however, that the scoring method chosen only has a serious effect on reliability if the test's participant's know which method they will be assessed by (Bachman, 1990:124), as knowing that exactly-correct answers are required reduces the occurrence of deliberately "creative" responses.

When presenting a fixed ratio cloze test as a measure of general language proficiency, content validity is demonstrated by "the representative sample of the linguistic features of the text" (Hughes, 2003:189) obtained by deleting every nth word. Cloze test correlations with other tests are affected by the chosen deletion rate, text variation, and the scoring method applied (Alderson, 1979:226), with SEMAC scoring marginally but consistently superior to EXACT scoring (Owen, et al., 1997:43). A question over the validity of cloze tests in general remains, as in some cases native speakers have performed less successfully than L2 speakers (Hughes, 2003:189). They may also lack face validity, being "perceived by many test takers as a highly artificial and 'untestlike' task" (Bachman, 1990:48).

The remainder of this article describes an experiment carried out using two different cloze tests based on the same text, and how the results of these varied.

3. TEST METHOD

3.1 Cloze test text

The text used to obtain the cloze tests in this experiment was a 162-word piece of writing produced by a previous year's student at the same level as those participating in the test. This text was described by the students' teacher as a fairly typical piece, which had been checked and corrected by the teacher before being submitted. It therefore represents a sample of the kind of writing that students at this level could reasonably be expected to process and produce themselves, having been taught all of the grammatical structures and the vast majority of the vocabulary contained within it. It was supposed that a cloze test utilising this text would therefore provide useful results, as most of the participants would be able to make reasonable attempts at understanding the text and restoring its missing words using the limited knowledge of language that they were expected (by their teacher) to possess. The original text can be seen in Appendix A.

3.2 Cloze test construction

It was decided that this experiment would utilise the fixed ratio deletion method of cloze construction, in the belief that this method would better produce tests capable of overall measuring language proficiency (rather than tests concentrating on particular parts of speech, or verb forms, for example). It was also decided that two different tests would be produced from the same text. This allowed comparison between the results of the two tests, and analysis of the balance of item types being tested in each. To produce each test, the first sentence was left unaltered, then every fifth word was deleted from the second sentence onward, only at different starting points (the fifth word in the sentence for Test 1, the second word for Test 2). This resulted in two cloze tests, each with 25 unique test items (missing words).

3.3 Test participants

The participants selected for this experiment were fifty Japanese junior high school students in their first year of formal English instruction, aged 11 and 12. Of the classes made available for participation, the two with the closest average scores on their most recent term test were chosen. These two classes also happened to have the same number of students present on the day of the cloze test's delivery. Although the level of English ability varied between individuals in the two classes, the English teacher that they shared described them as very similar in terms of ability, motivation and behaviour overall.

3.4 Test procedure

The two classes were each given one of the cloze tests. Students in the same class were all given the same test, in order to eliminate the possibility of colluding students reconstructing the complete original text by comparing tests 1 and 2. The test was conducted under normal "test conditions", in silence and with students instructed not to confer with others or consult other materials. These instructions, as well as an explanation that they were expected to attempt to restore the text to its original form by filling in the missing words as accurately as possible, were given in both English and Japanese. The students were then allowed as much time as they required to complete the test (which was between twenty and twenty-five minutes for each of the two classes). The participants were informed of the purpose of the test (and that it was the test itself, rather than them, being assessed) only after completing it, in order to maximise their motivation to complete it successfully and to the best of their ability.

3.5 Marking procedure

After collecting the fifty test papers, each was marked by a single marker and their totals were tallied. Initially, each paper was marked by the "exact answer only" scheme, which required no judgment beyond comparison with the original text (except in the case of misspellings, which were marked as correct as long as they unambiguously referred to the correct word in the correct form). Next, each paper was marked by the "semantically acceptable" scheme. Here, any word that could conceivably fit in a given gap was marked as correct, providing it did not create a logical conflict with other information provided in the text. Since acceptability for the answers on all of the test papers was judged by only a single marker, consistent judgment criteria could be followed. Under both methods of marking, correct (or acceptable) answers were simply given one point each, and incorrect (or unacceptable) answers were given none.

4. RESULTS

Table 1 shows that the two groups' recent term test scores were very similar, but that the group taking cloze Test 1 scored slightly higher overall by both the SEMAC and EXACT scoring methods.

Table 1 Mean scores for both groups in their recent term test and their cloze test

	Term test (out of 50)	Cloze SEMAC (out of 25)	Cloze EXACT (out of 25)
Test 1	30.67 SD = 13.71	8.63 SD = 6.44	6.37 SD = 4.82
Test 2	29.33 SD = 14.4	7.73 SD = 5.11	5.90 SD = 3.99

In Table 2, it can be seen that the EXACT scoring method resulted in a much smaller proportion of correct answers to content items than to function items. However, there was no clear pattern of function items being more successfully answered than content items in either of the tests when using a single scoring method.

Table 2 Mean scores for both groups on function word items and content word items

		SEMAC scoring	EXACT scoring
Test 1	**Function items (out of 10)**	3.80	3.53
	Content items (out of 15)	4.83	2.83
Test 2	**Function items (out of 13)**	3.87	3.57
	Content items (out of 12)	3.87	2.33

Correlations between the participants' scores when the SEMAC and EXACT methods were used can be seen in Table 3. The Pearson product-moment correlation coefficient measures the strength of the correlation between two sets of point scores. Spearman's ρ measures the strength of the correlation between two sets of rankings. With full sets of 25 items, these correlations are comparably very strong in both Tests 1 and 2. However, when content word-based items are isolated and scores for them are analysed, the PPM correlation is noticeably weaker (although the result can still be considered a reasonably strong correlation).

Table 3 Correlations between SEMAC scores and EXACT scores

	Pearson product-moment correlation coefficient			**Spearman's ρ**
	All items	Function items	Content items	
Test 1 SEMAC : EXACT	0.98	0.99	0.91	0.98
Test 2 SEMAC : EXACT	0.98	0.98	0.89	0.95

Correlations between participants' school term test scores and their cloze test scores can be seen in Table 4. There is little difference when different scoring methods are used. However, some variation can be seen in the correlations between term test scores and cloze test scores for different item types (correlations with function items being generally less strong that those with content items).

Table 4 Correlation between term test scores and cloze test scores

	Pearson product-moment correlation coefficient			**Spearman's ρ**
	All items	Function items	Content items	
Test 1 Term test : SEMAC	0.83	0.73	0.83	0.84
Test 1 Term test : EXACT	0.80	0.70	0.79	0.81
Test 2 Term test : SEMAC	0.66	0.55	0.68	0.58
Test 2 Term test : EXACT	0.65	0.56	0.68	0.59

Table 5 gives reliability estimates for the two cloze tests using each scoring method, and two different formulas. In nearly all cases, using either formula (and two different methods of splitting the data for Guttman's estimate), reliability estimates are slightly higher when using the SEMAC scoring method than when using the EXACT method.

Table 5 Reliability estimates for the two cloze tests using both scoring methods

		Guttman's split-half reliability estimate		**KR-20 reliability estimate**
		Odd-numbered items : Even-numbered items	Items 1-13 : Items 14-25	
Test 1	**SEMAC**	0.92	0.94	0.92
	EXACT	0.84	0.90	0.87
Test 2	**SEMAC**	0.89	0.69	0.87
	EXACT	0.89	0.67	0.83

5. DISCUSSION

5.1 Performance on the two tests

The difference in the performance of the two groups might well be explained by factors external to the test, since the two test groups consisted of different students, and the two tests were carried out in different rooms at different times. However, it is worthwhile considering the possible effects of the differences between the two tests themselves. Having different starting points for deletion of words, the two tests' items are necessarily different. Test 1 contained 50% more content word items than function word items, which could have been expected to make it the more difficult of the two tests (Test 2's items were almost equally balanced between content and function words). However, Test 1's scores were higher under both scoring methods. This can be perhaps be explained by the limited vocabulary of students at this level,

and the fact that the student who produced the text used in the test would have been exposed to similar content words as the students taking the test. Therefore, when required to produce a content word to fill a gap, the test participants would be drawing from the same limited vocabulary, and would be more likely to select a word that is not only semantically acceptable, but exactly correct. This might have had an equalising effect on the difficulties of content and function items in the test.

5.2 Validity of the two tests

An attempt at establishing concurrent validity for the two cloze tests was made by comparing students' cloze scores with their most recent school term test scores (as shown in Table 4). Term test scores correlated more highly with cloze Test 1 scores. This suggests that cloze Test 1, with its greater proportion of content word items, tests language proficiency in a more similar way to the term test than cloze Test 2 does. This is a difficult observation to explain, since the term test focused mainly on technical aspects of grammar (verb forms, article use, sentence structure), which might be expected to correlate more highly with a cloze test containing more function word items. Therefore, it might reasonably be supposed that cloze Test 1's items are simply easier than those in Test 2 (possibly due to them being more frequent in either the text or in the language as a whole).

5.3 Reliability of the two tests

The reliability estimates of both close tests shown in Table 5 indicate a generally acceptable level of reliability (certainly for a class-based test with no real consequences for its takers). Unfortunately, there are problems with each of the methods used to produce these figures. Guttman's split-half estimate operates under the assumption that items in the test are independent of each other (Bachman,

1990:174), which is not the case in a cloze test (Owen, et al., 1997:42). Therefore it should be assumed that the actual reliability of each test is lower than that indicated by the Guttman figures in the table. The KR-20 formula also assumes item independence, but additionally assumes that each item is equivalent in its level of difficulty (Bachman, 1990:177), which cannot be guaranteed in a close test. Therefore the KR-20 reliability figures cannot themselves be considered reliable.

5.4 Performance under the two scoring methods

Table 1 clearly shows that students' cloze scores were higher when the SEMAC scoring method was used than with the EXACT method. Table 2 shows that scoring method in fact had very little effect on scores for function word items, but a much greater effect on those for content word items. This can be explained by the greater "openness" of content items, in that there is a greater number of possibly-correct answers for a content item than there are for a function item. Therefore, students' answers to content items show greater variation; fewer of them will be exactly correct when compared to the original passage, but more of them will be nevertheless acceptable in the given context.

Correlations between SEMAC scores and EXACT scores were high in all cases and by all methods of calculation. This would seem to support the claim that either scoring method might be used, and that the less time-consuming EXACT method would therefore be preferable (especially to a teacher scoring a large number of class tests). However, greater discrimination between test-takers at different levels is possible using the SEMAC method, especially as shown above when test items are more often content words than function words.

5.5 Validity under the two scoring methods

As previous research has shown, the SEMAC scoring method seemed to result in very slightly higher concurrent validity coefficients than the EXACT method. Both the PPM formula for correlating absolute scores and Spearman's formula for correlating rankings demonstrated this tendency. Again, this indicates that the SEMAC method may be more appropriate if cloze tests were to be used more often with these students, as it seems to provide the more similar measure of proficiency to that tested by the school's term test. Since students tend to be taught specifically for the purpose of passing tests of this kind, a SEMAC-scored cloze test might provide an alternate testing method.

5.6 Reliability under the two scoring methods

The reliability estimates used, despite their own inherent unreliability (discussed above), indicate that the SEMAC method is the (slightly) more reliable of the two. This is less clearly the case in the results of Test 2, where the lack of independence between items may have affected the two Guttman estimates provided.

6. CONCLUSIONS

Although it may not be desirable for cloze tests to become a regular feature of the classroom (or even a more commonly-used testing tool), it seems that the SEMAC scoring method is the most appropriate should such tests be carried out with students at this level. Although students with limited vocabulary are less likely to produce a great variety of responses to "open" items, there is little justification for penalising them for filling gaps with words that fit but do not match an unseen original text.

The two tests created for this study were each presented to a different set of students. This meant that there were variables present beyond those in the test

itself, such as the different language proficiencies of the two classes, different class atmosphere during the test (it is possible that one class simply took the test more seriously and tried harder than the other), instructions explained slightly differently to the two groups, and so on. Although the two tests could not have been presented to a single group of students (at least without a considerable period of time between the two, during which they would hopefully forget the test's content), it would be desirable to find a more effective way to compare two tests created from a single text with different words deleted. This would give greater insight into the effect that various deletions have on students' ability to restore them, without the need to account for the confounding variables described above.

A student-generated text was used to create the cloze tests. It was hoped this text would be both level-appropriate for the test-takers (students at the same point in their language learning) and deal with appropriate topics. The fixed ratio deletion method used helped to ensure coverage of different language features. These techniques meant that reliability data is hard to calculate, as there was no way to ensure the equivalence and independence of test items. Future work might attempt to split the tests into more equivalent halves, or to choose words for deletion more selectively. This would allow more stable reliability estimates, and possibly make clearer any differences between the two scoring methods.

REFERENCES

Abraham, R. G. and Chapelle, C. A. (1992). The Meaning of Cloze Test Scores: An Item Difficulty Perspective. **The Modern Language Journal**, 76/4, 468-479.

Aitken, K. G. (1977). Using Cloze Procedure as an Overall Language Proficiency Test. **TESOL Quarterly**, 11/1, 59-67.

Alderson, J. C. (1979). The Cloze Procedure and Proficiency in English as a Foreign

Language. **TESOL Quarterly**, 13/2, 219-227.

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.

Bachman, L. F. (1990). **Fundamental Considerations in Language Testing**. Oxford University Press.

Bormuth, J. R. (1969). Factor Validity of Cloze Tests as Measures of Reading Comprehension Ability. **Reading Research Quarterly**, 4/3, 358-365.

Brown, H. D. (2000). **Principles of Language Learning and Teaching** (4th edition). Longman.

Brown, J. D. (1980). Relative Merits of Four Methods of Scoring Cloze Tests. **The Modern Language Journal**, 64/3, 311-317.

Hughes, A. (2003). **Testing for Language Teachers** (2nd edition). Cambridge University Press

Jonz, J. (1990). Another Turn in the Conversation: What Does Cloze Measure? **TESOL Quarterly**, 24/1, 61-83.

Kobayashi, M. (2002). Cloze Tests Revisited: Exploring Item Characteristics with Special Attention to Scoring Methods. **The Modern Language Journal**, 86/4, 571-586.

Oller, J. W. (1972). Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language. **The Modern Language Journal**, 56/3, 151-158.

Oller, J. W. and Inal, N. (1971). A Cloze Test of English Prepositions. **TESOL Quarterly**, 5/4, 315-326.

Owen, C., Rees, J., Wisener, S., and Crompton, P. (1997). **Testing**. Centre for

English Language Studies, University of Birmingham.

Richards, J. and Schmidt, R. (2002). **Longman Dictionary of Language Teaching and Applied Linguistics**. 3rd ed. Longman.

Stubbs, J. B. and Tucker, G. R. (1974). The Cloze Test as a Measure of English Proficiency. **The Modern Language Journal**, 58/5, 239-241.

APPENDIX A – Text used to produce cloze tests

Hello, my name is Emiko. I come from Osaka, but now I live in Shizuoka with my family. They are my mother, my father, and my two sisters. My mother and father work in a shop. My sisters are elementary school students, but I am a junior high school student. My class is 1-6, and my teacher's name is Mr. Watanabe. He is an English teacher. Anna teaches English at my school too. She is from Australia. I like English, but my favorite subject is P.E..

I like sports very much. I play tennis and soccer at school and with my friends. I'm a member of the volleyball club, so I practice volleyball on Saturdays and Sundays. I don't like swimming very much.

I arrived in Shizuoka three years ago. I have lots of friends here now. I like Shizuoka because it has lots of nice mountains and beaches. It also has shops, schools and libraries, so it is an interesting place.

APPENDIX B – Cloze test 1

Hello, my name is Emiko. I come from Osaka, but now I live in Shizuoka with my family. They are my mother, (_____) father, and my two (_____). My mother and father (_____) in a shop. My (_____) are elementary school

students, (_____) I am a junior (_____) school student. My class (_____) 1-6, and my teacher's (_____) is Mr. Watanabe. He (_____) an English teacher. Anna (_____) English at my school (_____). She is from Australia. (_____) like English, but my (_____) subject is P.E..

I (_____) sports very much. I (_____) tennis and soccer at (_____) and with my friends. (_____) a member of the (_____) club, so I practice (_____) on Saturdays and Sundays. (_____) don't like swimming very (_____).

I arrived in Shizuoka (_____) years ago. I have (_____) of friends here now. (_____) like Shizuoka because it (_____) lots of nice mountains and beaches. It also has shops, schools and libraries, so it is an interesting place.

APPENDIX C – Cloze test 2

Hello, my name is Emiko. I come from Osaka, but now I live in Shizuoka with my family. They (_____) my mother, my father, (_____) my two sisters. My (_____) and father work in (_____) shop. My sisters are (_____) school students, but I (_____) a junior high school (_____). My class is 1-6, (_____) my teacher's name is (_____) Watanabe. He is an (_____) teacher. Anna teaches English (_____) my school too. She (_____) from Australia. I like (_____), but my favorite subject (_____) P.E..

I like sports (_____) much. I play tennis (_____) soccer at school and (_____) my friends. I'm a (_____) of the volleyball club, (_____) I practice volleyball on (_____) and Sundays. I don't (_____) swimming very much.

I (_____) in Shizuoka three years (_____). I have lots of (_____) here now. I like (_____) because it has lots of nice mountains and beaches. It also has shops, schools and libraries, so it is an interesting place.