贅詞減量的語言測驗：
克漏字與 C 測驗的重新檢驗

林文鶯[*]　袁曉青[**]　馮和平[***]

本研究的主要目的：(1)探究採用由 Farhady 與 Keramatic（1996）兩位學者所提出的由文本決定挖空方法（text-driven method）的克漏字測驗，與標準克漏字測驗（standard cloze test）比較起來，前者是否會產生較好的測量特質（psychometric properties）；(2)比較克漏字測驗和 C-test 的測量特質；(3)探究採用不同的挖空比例是否會導致考生的考試成績表現高低不同；(4)驗證「考生不會因採用不同文本而產生不同的考試成績表現」（test-takers’ performance invariance across different texts）這個假定（assumption）是否適用於克漏字測驗和 C-test。本研究的受試者是來自北臺灣一所大學的二百三十七位大一學生。本研究的結果顯示從信、效度的層面來看，由文本決定挖空方法的三個克漏字測驗版本及 C-test，並沒有比標準克漏字測驗版本來得好。針對不同的挖空比例是否會導致考生的考試成績表現高低不同這個研究問題，本研究似乎未能得到具有結論性的證據。本研究也未得到強而有力的證據來支持「考生不會因採用不同文本而產生不同的考試成績表現」這個假定。

關鍵字：贅詞減量的語言測驗、克漏字、挖空比例

* 作者現職：臺北市立教育大學英語教學系副教授

**作者現職：銘傳大學語言中心講師

***作者現職：國立臺灣師範大學英語系副教授

1. Theoretical Background

Language reduced redundancy (hereafter referred to as LRR) approach to language testing, proposed by Spolsky, Bengt, Sako, and Aterburn (1968), has been a basis for developing numerous major test procedures, such as cloze test, C-test, dictation, etc. According to Spolsky et al., the phenomenon of redundancy utilization occurs in everyday language use. For instance, in noisy surroundings, a person often has to guess at the words s/he cannot hear by relying on the whole conversation. That is, predicting and supplying missing linguistic information in a message is a normal activity in daily life. Hence, Spolsky et al. argued that knowing a language involves the ability to understand an incomplete or distorted message and make educated guesses about a certain percentage of the missing information. They also contended that as a learner's proficiency in a language improves, s/he will be able to make more successful use of the redundancy inherent in the language and obtain a higher score on a LRR test. In a nutshell, the key rationale of the LRR approach is that to test a person's command of a language is to evaluate his/her ability to make use of the redundancies inherent in the language by asking him/her to guess the omitted linguistic elements.

To operationalize this rationale, LRR tests are generally implemented by presenting an examinee with a piece of mutilated text and asking him/her to restore the text. Technically, LRR tests are based on two steps of random sampling. The first step of random sampling occurs when test constructors select a text for test construction. As reviewed by Klein-Braley (1997), a text used for test construction under the LRR approach basically functions as a sample of the language. That is, a text used for LRR tests theoretically should be the result of a random sampling procedure and thus should be interchangeable with any other text. Hence, authentic texts are often recommended for use for LRR tests to approximate random sampling of texts. The second random sampling takes place when LRR tests incorporate random noise by using a random (or pseudo-random) deletion technique for test construction. The elements randomly deleted from the text function in the same way as noise randomly occurs in a communication system and are considered as a random sample of all the elements in the text. Therefore, with random sampling as the cornerstone, LRR tests aim at obtaining a random sample of an examinee's performance under a test setting where random noise is deliberately included. That is, given a text with some of its elements randomly deleted, an examinee is required to exhibit a random sample of his/her language ability in LRR tests. How s/he performs in LRR tests under controlled condition of “random noise” is believed to provide evidence of his/her language proficiency.

Cloze Tests

Of the numerous LRR-based testing procedures, cloze test, according to Klein-Braley (1997), is the most important and best-known operationalization of the LRR principle. Developed originally by Taylor (1953) as a test for measuring the readability of native English speakers’ texts, cloze test was investigated in the 1960s by a number of researchers (Bormuth, 1965, 1967; Crawford, 1970; Gallant, 1965; Ruddell, 1964) as a potential measure of native English learners’ reading proficiency. In the 1970s, another line of research (e.g., Alderson, 1979; Oller, 1973) investigated the effectiveness of cloze test as a measure of overall ESL/EFL proficiency. According to Hinofotis (1980), the word “cloze” comes from the concept of closure used in Gestalt Psychology and refers to the ability to fill in the gaps in an incomplete pattern. The LRR principle is highly involved in cloze test in the sense that the test reduces natural linguistic redundancies and requires examinees to utilize organizational constraints to infer meaning and fill in the blanks.

Among various forms of cloze test, standard or fixed ratio form has been extensively investigated from the methodological perspective. Standard cloze form consists of a text, from which a word is deleted after every certain number of words according to an arbitrary and fixed ratio procedure. For example, every seventh or tenth word is deleted after one or two sentences of unbroken text. The examinee is required to supply the missing words by inferring from the context. This systematic deletion of words used in standard cloze was suggested by Taylor (1953) as an efficient way to approximate random deletion using random number table. Among numerous researchers who investigated standard cloze, Oller (1973; 1979) was most famous for actively popularizing it as a highly effective way of measuring a learner’s overall second/foreign language proficiency. He argued that the actual text used and the actual deletion employed for test construction are irrelevant because tests that use different texts with different levels of difficulty or tests that employ different deletion rates will still rank examinees in the same order. In other words, Oller claimed that the assumption about the test-takers’ performance invariance across the standard cloze with different texts or different deletion rates is tenable.

However, research findings on this assumption are not conclusive. For example, Klein-Braley (1981) and Zarrabi (1988) demonstrated that different texts using the same deletion rate will result in different reliability and different correlations with the criterion measures. This suggests that selecting different texts may produce different tests, each of which measures certain aspects more effectively than other aspects. As for the robustness of cloze test to different deletion rates, Alderson (1979) found that deletion rate is an important factor affecting the results of standard cloze. In particular, he found that a text can produce quite different tests depending on whether, say, every seventh rather than every tenth word is deleted. He claimed that much of the discrepancy in examinee performance on cloze tests may be due to the deletion of different words through different deletion rates. Therefore, he refuted the principle of randomness required by the LRR approach and favored the view that the deletion should be based on "a theory of the nature of language and language processing." (p. 226)

Farhady and Keramati (1996) voiced the same opinion in their study to support the suggestion of Weaver and Kingston (1963) and Ohnmacht, Weaver, and Kohler (1970) that the deletion rate of a cloze test should be based on the number of linguistic and discourse structures of a text (i.e., text-driven deletion method) and not on an arbitrary number (e.g., 5, 7, or 9). In their study, standard cloze and eight other different forms of cloze test, all based on a single text of 337 words about telepathy, were constructed and administered randomly to 403 Iranian students at the University for Teacher Education. In contrast to the standard cloze in which the deletion rate was set arbitrarily at 7, the deletion rate of the other eight cloze forms was set on the basis of the text's number of sentences, T-units, dependent clauses, independent clauses, noun phrases, verb phrases, adjectival phrases, or cohesive ties. Specifically, the deletion rate for each of the eight forms was determined by Farhady and Keramati's formula, which took into account the number of the linguistic structures of the text. Their results showed that of the nine cloze forms, the one in which the deletion rate was based on the number of existing noun phrases of the text produced the best psychometric properties, in terms of both criterion-related validity and reliability estimates. The second best was the one in which the deletion rate was based on the number of verb phrases. Both forms, according to Halliday and Hassan (1976) and Halliday (1985), were based on the number of linguistic structures below the clause level of the text. On the other hand, Farhady and Keramati found that again in terms of criterion-related validity and reliability estimates, the standard cloze with arbitrary, fixed-ratio method generally failed to produce a better test than the close forms that were based on their text-driven deletion method. Hence, in their conclusion, Farhady and Keramati warned against the use of the former but favored the use of the latter. In particular, they strongly advocated the use of the cloze forms that set the deletion rate according to the number of linguistic structures at below the clause level.

However, a closer look at their study points to a need for careful interpretation of their results. In their study, the adjusted reliability coefficient for the standard cloze was 0.76, which ranked fourth and was only slightly smaller than those (0.77, 0.79, and 0.84) of the three forms that were based on the text-driven deletion method. Similarly, the criterion-related validity coefficient of the standard cloze with vocabulary criterion measure was 0.52, which also ranked fourth and again was only slightly smaller than those (0.61, 0.59 and 0.56) of the three forms that were based on text-driven deletion method. In fact, the criterion-related validity coefficient of the standard cloze with structure criterion measure was 0.69, which was as high as that of the close form that was based on noun phrases of the text. Hence, their conclusion that cloze forms based on text-driven method are superior to the standard cloze seems premature. Moreover, their study was based on a single text. As Gamarra and Jonz (1987), and Jonz (1989) pointed out, type of the text was one of the factors affecting the validity of the cloze test. With this recognition, Farhady and Keramati (1996) indicated that their conclusion about what basis the deletion rate of a cloze test should be on cannot be firmly drawn unless further research with different texts confirms their findings.

C-Tests

Growing out of the dissatisfaction with unpredictably non-equivalent results caused by different deletion techniques used for constructing cloze test, C-test was proposed by Raatz and Klein-Braley (1981) as an alternative to cloze test. The “C” in C-test was chosen as an abbreviation of the word “cloze” to emphasize the relationship between C-test and cloze test. Also a representative of the LRR family, C-test, as pointed out by Raatz and Klein-Braley, was developed not only to retain the positive aspect of cloze test (i.e., its capacity to tap an examinee’s ability to process discourse and to predict from context with reduced redundancy) but also to correct the major technical defect of cloze test (i.e., the failure of its deletion technique to ensure a random sampling, which is crucial for LRR tests). Unlike cloze test in which deletion is performed at the text level, C-test was designed to achieve random sampling by performing deletion at the word level. That is, only parts of a word, rather than a whole word, are removed in C-test. Specifically, in C-test, the second half of every other word is deleted, leaving the first sentence of the text intact. If a word has an odd number of letters, then the larger “half” is deleted. If a word has only one letter (e.g., “I” and “a”), then this one-letter word is ignored in the counting. By deleting at the word level, C-test, claimed by Raatz and Klein-Braley, can produce a more representative sample of the elements of the text than cloze test.

However, like cloze test, C-test has been subject to intense debate over the past 20 years. On one side are its advocates who considered it as a theoretically and empirically valid measure of general language proficiency and claimed it as a good substitute for cloze test (Dornyei and Katona, 1992; Grotjahn, 1986, 1987; Klein-Braley, 1997). For instance, in his study comparing the empirical performance of C-test and other LRR-based tests (such as standard cloze, multiple-choice cloze, and cloze-elide), Klein-Braley (1997) found that C-test is the most economical and reliable procedure and has the highest empirical validity. On the other side are those who argued against the superiority of C-test but considered it as an instrument for measuring examinee’s ability to utilize the knowledge of word structure rather than for measuring their ability to process discourse for general proficiency (Carroll, 1987; Cohen, Segal, and Weiss, 1984; Hughes, 1989; Jafarpur, 1995, 1996; Weir, 1988). For example, Jafarpur (1995; 1996) expressed his skepticism about Klein-Braley’s claim about the superiority of C-test over cloze test. In addition to a lack of face validity of C-test, Jafarpur found that similar to those for cloze test, the underlying assumptions of random sampling of the basic elements of a text were not tenable for C-test since various deletion ratios and deletion starts produced different C-tests. Therefore, in light of the differences in viewpoint about the validity and superiority of C-test over cloze test, a study to re-examine the empirical performance of these two tests is definitely in order.

2. Research Questions

Based on the above review of relevant literature, this study is conducted, partly replicating and partly extending the study of Farhady and Keramati (1996), in an attempt to answer the following research questions: (1)Would the cloze form that is based on text-driven deletion method produce better psychometric properties than the standard cloze form? (2) Is C-test superior to various forms of cloze test in terms of reliability and validity? (3) Will different deletion rates lead to differences in test-takers’ performance? (4) Does the assumption of test-takers’ performance invariance across different texts hold for both C-test and various forms of cloze test?

3. Method

Subjects

A total of 237 (198 female and 39 male) sophomore students majoring in Applied English at one private university in the northern Taiwan were the subjects of the study. Like other university students in Taiwan, they all received at least seven to eight years of formal English instruction prior to their college education. Overall, they received 3-5 hours of English instruction a week in Junior high schools and 4-8 hours in senior high schools. They were taught mainly under grammar-translation approach, communicative teaching approach, or a combination of both. As most students admitted to the department of the university fell within an average-grade category based on their English performance in the Joint College Entrance Examination, their English proficiency should not be different substantially.

Instruments

Two authentic texts were selected as the basis for constructing four different forms of cloze test and one form of C-test, which were the major instruments in this study. For the purpose of examining the empirical validity for the different forms of the tests, three criterion measures were used. The texts, the different forms of the tests, and the criterion measures are described as follows:

Texts

Numerous proponents of LRR tests suggest that authentic texts (usually claimed as representing genuine samples of languages in use) should be used to approximate random sampling, which is required by LRR tests (see, for example, Klein-Braley, 1997). An authentic text, in its strictest definition, refers to the text that is written for native speakers, rather than specifically for second or foreign language learners. In an authentic text, no modification or adaptation is made for the purpose of teaching second or foreign language learners any particular target sentence patterns or vocabulary items. Based on this definition, six authentic texts were initially selected for this study. Of the six texts, three were in narrative mode, ranging in length from 451 to 536 words; the other three were in persuasive mode, ranging in length from 313 to 418 words. Each of the six texts was read by 11 students who had similar backgrounds to those of student subjects in the formal study. Since deleting a number of words from a text in LRR tests will considerably decrease its readability (thus increasing test takers' reading difficulty), choosing a text that is too difficult will cause frustration for test takers. Madsen (1983) therefore suggested that LRR test constructors should choose a passage that test takers can read with little or no difficulty at all. Following his suggestion, the present study required the 11 students to choose the two easiest texts, one from the three narrative texts and the other from the three persuasive texts. Of the three narrative texts, the one with the story "Sylvester and the Magic Pebble" was chosen by most students to be the easiest (7 out of 11) and the most interesting (7 out of 11) text (Appendix A). Of the three persuasive texts, the one with the title "Do you want to be Wise? Rich? Famous?" was chosen by most students to be the easiest (6 out of 11) and the most interesting (7 out of 11) text (Appendix B). Accordingly, these two texts were used for constructing the different test forms for the study. The two texts are described briefly as follows:

Text I. The first text is an extraction from the story "Sylvester and the Magic Pebble" by Steig (1969). The rhetorical mode of Text I is basically narrative. The text, containing 483 words and 11 paragraphs, is estimated to have the Flesch-Kincaid grade level of 6.1. That is, in terms of reading difficulty, it is suitable for sixth graders in the United States. Text I is about how a little donkey accidentally uses a pebble to make itself become a rock to avoid falling prey to a lion.

Text II. The second text is an essay taken word for word from the article "Do you want to be Wise? Rich? Famous?" written by Van Doren and compiled in "A Reader for Writers" (Lee, 1988). Text II is a persuasive/argumentative essay. The text, containing 313 words and six paragraphs, is estimated to have the Flesch-Kincaid grade level of 3.5. That is, it is suitable for third to fourth graders in the United States. The main idea of Text II is to persuade people to realize that one needs to pay for whatever he/she desires to possess.

One thing worth mentioning here is that, in terms of their Flesch-Kincaid grade levels, the two chosen texts seemed appropriate for third to sixth graders, rather than for the subjects of the present study, who were sophomores and had at least seven to eight years of formal English training. However, as English is a foreign language in Taiwan, where students generally have limited English exposure outside the classroom, the subjects' reading ability tended to be much poorer than that of the English-native-speaking sixth to eighth graders. In addition, as mentioned earlier, Madsen (1983) highly recommended choosing a LRR passage that test takers can read with little or no difficulty. Therefore, the two texts (with the Flesch-Kincaid grade levels of 6.1 and 3.5) were chosen so as to better fit the subjects' reading ability.

Another point, also related to the Flesch-Kincaid grade level, deserves mentioning as well. Text I and Text II appeared to differ in terms of their readability level (as measured by the grade level). Specifically, the Flesch-Kincaid grade level of 6.1 for Text I is relatively higher than that of 3.5 for Text II. That is, the difference of 2.6 in the grade level appeared to suggest that Text I may be more difficult in readability than Text II. However, note that the Flesch-Kincaid grade level, like most readability indexes, fails to take into account the text's organization and readers' familiarity with its organization. It is commonly held that people in almost every culture are all quite familiar with narratives, as people tend to have had plenty of exposure to stories ever since they were small kids. As such, people in general are somewhat familiar with narratives, the contents of which are usually organized in time sequence. Hence, they often encounter fewer difficulties when reading narratives than when reading other types of texts, the contents of which are usually not organized in time sequence. In other words, given the same Flesch-Kincaid grade level for two texts with two different rhetorical modes (one of which is narrative), the sequential nature of the narrative generally tends to help facilitate readers' comprehension and make them feel that it is less difficult in readability than the other text with a different rhetorical mode. Hence, to ensure that the subjects would feel the same in terms of reading difficulty for the two texts, Text I was deliberately chosen to have a higher Flesch-Kincaid grade level than Text II.

Five Different Test Forms

Based on the two texts, four different forms of cloze test and one form of C-test were constructed. For all test forms, the first two sentences of the texts were left intact to give the subjects more information to understand the overall meaning of the text and to help them become familiar with the style of the texts.

The first three forms of cloze test had their deletion rates set on the basis of the number of dependent clauses (Form A), noun phrases (Form B), and verb phrases (Form C). Following Farhady and Kermati (1996), the deletion rates for the three forms were determined by leaving the first two sentences intact and dividing the remaining number of words by the number of specified structures. For example, for

Text I, there is a total of 475 words and a total of 22 dependent clauses. The number of words for the first two sentences is 25. Thus, for Form A and Text I, the deletion rate of 20 was obtained by subtracting 25 from 475 and divided by 22. Based on the results of these calculations, the fixed ratio method was then employed and the three forms of cloze test constructed. Table 1 presents the structures on which the three forms of cloze test were based. It provides the number of deletions and the deletion rate for each form. These three forms were specifically chosen for this study because the results of Farhady and Kermati (1996) indicated that the cloze test that is based on the number of dependent clauses has the worst reliability estimates, whereas the cloze test that is based on the number of noun or verb phrases has the best reliability estimates and/or criterion-related validity.

Form D and Form E were constructed without considering the linguistic structure of the texts. Specifically, Form D was the standard cloze in which the deletion rate was arbitrarily set at 7 and Form E is C-test in which the second half of every other word was deleted.

Table 1 The Number and Rate of Deletion in Each Test Form

Text	Test form	Linguistic & discourse structure	No. of deletion	Deletion rate
I	A	Dependent clauses	22	20
	B	Noun phrases	90	5
	C	Verb phrases	75	6
	D	Standard cloze	64	7
	E	C-test	214	2
II	A	Dependent clauses	12	23
	B	Noun phrases	70	4
	C	Verb phrases	47	6
	D	Standard cloze	40	7
	E	C-test	136	2

Three Criterion Measures

Three criterion measures -- structure and written expression, vocabulary, and reading comprehension -- were applied in this study to validate the different test forms. Set up in multiple-choice format with four options for each item, the three measures were subtests from a sample test of TOEFL, which is well-known for assessing the general English proficiency of people whose native language is not English. Furthermore, the Cronbach's reliability coefficients of the three measures were estimated at 0.76 for the structure subtest, 0.79 for the vocabulary subtest, and 0.67 for the reading subtest. These estimated coefficients were smaller than those that are normally reported in TOEFL manual. This discrepancy was not unexpected, considering the fact that the study used much smaller sample size and thus found much narrower range in students' English proficiency than pilot studies conducted by Educational Testing Service to validate TOEFL.

Procedures

The five test forms and the three criterion measures were administered to the student subjects over a month period (from mid- March to mid April) during their enrollment in Rhetorical Writing in their fourth semester at the university. Specifically, the five test forms were distributed randomly among the subjects such that those taking Form A were classified as Group A, those taking Form B as Group B, those taking Form C as Group C, those taking Form D as Group D, and those taking Form E as Group E. To control for order effect, half of the subjects were arranged to take Text I first and then Text II, and the other half to take Text II first and then Text I. The time allocated was 30 minutes for each text and 75 minutes for the criterion measures. The time interval between the test form and the criterion measures was about one or two weeks. The five test forms were graded using the exact word scoring method because it was found to be quite reliable and practical (Bailey, 1998). Furthermore, as the total number of deletions (or items) is different for each of the five test forms and between the two texts, the number of correct words restored for each test form and each text was transformed into percentage correct score multiplied by 100 for ease of comparison.

4. Results and Discussions

The basic statistics of the three criterion measures for the five groups of test form are presented in Table 2. Although there were some slight differences among the five groups in means and standard deviations of the three measures, results from the one-way ANOVA (analysis of variance) with groups (test forms) as the between-subject factor indicated that these differences were not statistically significant, with *F*(4,232) = 0.858 and *p* = 0.49 for the structure subtest, *F*(4,232) = 1.227 and *p* = 0.30 for the vocabulary subtest, and *F*(4,232) = 0.675 and *p* = 0.61 for the reading subtest. Thus, the random assignment of the five test forms among the subjects ensured that the five groups started out with roughly equal level of general English proficiency. Hence, any difference in mean scores of the five test forms should not be attributed to the preexisting difference in the level of general English proficiency among the five groups.

Table 2 Means and Standard Deviations (SD) of the Three Criterion Measures for the Five Groups

Treatment	No. of	Structure subtest		Vocabulary subtest		Reading subtest	
Groups	Students	Mean	SD	Mean	SD	Mean	SD
A	49	75.20	12.06	59.86	16.65	71.77	11.57
B	49	75.26	11.93	59.25	16.99	71.77	14.11
C	41	71.59	11.08	53.17	14.37	69.27	11.49
D	54	73.75	12.52	58.21	17.02	68.52	13.33
E	44	75.45	9.86	56.14	15.83	70.98	12.99

The reliability, one psychometric property, was compared across the five test forms using the Cronbach’s reliability coefficients. The results are presented in Table 3. The reliability coefficients of the five test forms ranged from a low of 0.43 (Form A) to a high of 0.96 (Form E) for Text I and from a low of 0.55 (Form A) to a high of 0.93 (Form E) for Text II. Since the reliability coefficient of a test form is affected by the number of items, the results for both texts are not unexpected. That is, Form A, which contained the smallest number of items, should produce the lowest reliability estimate, and Form E, which contained the largest number of items, should produce the highest reliability estimate.

Table 3 Reliability and Validity Coefficients of the Five Test Forms

Text	Test form	Reliability coefficient	Adjusted reliability coefficient	Validity coefficient (level of agreement %)		
				Structure subtest	Vocabulary subtest	Reading subtest
I	A	0.43	0.88	0.68 (46%)	0.69 (48%)	0.71 (50%)
	B	0.91	0.96	0.73 (53%)	0.61 (37%)	0.71 (50%)
	C	0.86	0.92	0.52 (27%)	0.58 (34%)	0.26 (7%)
	D	0.77	0.95	0.75 (56%)	0.59 (35%)	0.77 (59%)
	E	0.96	0.96	0.31 (10%)	0.34 (12%)	0.68 (46%)
II	A	0.55	0.93	0.77 (59%)	0.67 (45%)	0.69 (48%)
	B	0.87	0.93	0.71 (50%)	0.52 (27%)	0.64 (41%)
	C	0.65	0.84	0.79 (62%)	0.58 (34%)	0.33 (11%)
	D	0.71	0.89	0.61 (37%)	0.48 (23%)	0.82 (67%)
	E	0.93	0.93	0.30 (9%)	0.37 (14%)	0.49 (24%)

For the purpose of examining what the reliability of the four cloze forms (i.e., Forms A, B, C, and D) will become if their items increase to 214 for Text I and to 136 for Text II (as in Form E), the Spearman Brown Prophecy formula was used to calculate the adjusted reliability coefficients. The results in Table 3 indicated that all the five test forms were quite reliable for both texts and that the differences in magnitude of the adjusted reliability coefficients among the five test forms were, in general, quite trivial and insignificant. For example, for Text I, the most reliable test forms were Form B and Form E, both estimated at 0.96. They were followed by Form D, Form C, and Form A, estimated respectively at 0.95, 0.92, and 0.88. Similarly, for Text II, the reliability coefficients for Form A, Form B, and Form E were all estimated at 0.93, followed by Form D at 0.89 and Form C at 0.84.

However, a close examination of these adjusted reliability estimates revealed several interesting findings. For example, a comparison of the adjusted reliability estimates among the four cloze forms showed that Form B consistently produced highest reliability estimates for both texts, even though the superiority of Form B was slight and insignificant. Its slight superiority in the adjusted reliability estimates over the other test forms was also found in Farhady and Keramatic's study. As for Form A, its inferiority in the adjusted reliability estimate (0.39) found by Farhady and Keramatic was not consistently observed for the two texts in this study. As shown in Table 3, although the adjusted reliability estimate (0.88) of Form A was the lowest for Text I, yet for Text II, the adjusted reliability estimate (0.93) of Form A was as high as those of Form B and Form E. Furthermore, Form D (i.e., standard cloze form) did not necessarily produce lower adjusted reliability estimates than those of the other three cloze forms that were based on the text-driven deletion method. In particular, for both texts, its adjusted reliability estimates (0.95 for Text I and 0.89 for Text II) were consistently higher than those (0.92 for Text I and 0.84 for Text II) for Form C. Interestingly, for both texts, the adjusted reliability estimates of Form E (C-test) were as high as those of Form B. Both C-test and Form B tended to produce slightly better and relatively more stable reliability estimates.

A validity coefficient, another psychometric property, for each of the five test forms was also calculated. This coefficient was based on the Pearson product-moment correlation coefficients between the subjects' scores on the test forms and each of the three criterion measures. Because the validity coefficient is usually affected by the unreliability of its criterion measures, it is often suggested to be corrected for attenuation (Henning, 1987). The corrected validity coefficients for the test forms are shown in Table 3. These coefficients ranged from 0.26 to 0.77 for Text I and from 0.30 to 0.82 for Text II.

According to Hughes (2003), each of these corrected validity coefficients is better interpreted in terms of the level of agreement between each of the five test forms and each of the three criterion subtests. The levels of agreement can be computed by squaring the corrected validity coefficients and are presented in Table 3. The levels of agreement ranged from 10% to 59% for Text I and from 9% to 67% for Text II.

A close study of the different levels of agreement for the four cloze forms showed that their levels of agreement with the three criterion measures, in general, were not strikingly different across the four cloze forms. Interestingly, Form B, which was based on the number of noun phrases, did not have the highest criterion-related validity coefficients, as claimed by Farhady and Keramati (1996). Likewise, Form C, which was also based on the number of linguistic structures (verb phrases) below the clause level of the text, did not produce higher levels of agreement with all of the three criterion measures than the other cloze forms. Another interesting point to note was that Form D, where deletion rate is set arbitrarily, was not necessarily inferior to the other three cloze forms that are based on text-driven method. In fact, Form D had the highest levels of agreement (59% for Text I and 67% for Text II) with the reading criterion measure. Similarly, Form A, which was based on the number of dependent clauses, had the highest levels of agreement (48% for Text I and 45% for Text II) with the vocabulary criterion measure. The high level of agreement with the vocabulary criterion measure for both texts was not found by Farhady and Keramati.

As for the comparison of levels of agreement between the four cloze forms and C-test showed that for both texts C-test (Form E) appeared to tap abilities that are different from the four cloze test forms. Specifically, C-test had the lowest level of agreement with both the structure subtest (10% for Text I and 9% for Text II) and the vocabulary subtest (12% for Text I and 14% for Text II). The considerably low level of agreement (i.e., 12% and 14%) found in this study between C-test and vocabulary subtest was in sharp contrast with the high level of agreement (74%) found by Chapelle and Abraham (1990) between their C-test and their multiple-choice vocabulary test.

To examine whether different deletion rates lead to difference in test-takers' performance, the mean percentage correct scores (then multiplied by 100) of the four cloze forms for the four groups were calculated and presented in Table 4. The results revealed that no obvious difference existed for Text I but substantial difference existed for Text II. Specifically, Group C performed better than Group D and Group B, which in turn outperformed Group A. This result was certainly unexpected because Form A, which had the smallest number of blanks, should be easier than Forms B, C, and D, which had three to five times more number of blanks.

Table 4 Means and Standard Deviations (SD) of the Form Scores

Text	Test form	No. of students	Mean	SD
I	A	49	43.51	11.48
	B	49	42.68	13.50
	C	41	38.93	10.88
	D	54	43.46	9.60
	E	44	68.73	12.69
II	A	49	24.83	16.09
	B	49	40.47	12.17
	C	41	48.73	9.47
	D	54	40.74	11.68
	E	44	68.97	11.19

The above findings in mean performance difference among Groups A, B, C, and D were further confirmed by using the following statistical analyses. Specifically, results from one-way ANOVA using the test forms as the between-subject independent variable confirmed that there was no significant difference in mean performance among the four groups for Text I, with $F(3,189) = 1.57$ and $p = 0.20$. On the other hand, there was significant difference in mean performance among the four groups for Text II, with $F(3,189) = 29.06$ and $p = 0.00$. Furthermore, the results of the post hoc Scheffe procedure indicated that for Text II, significant differences in mean performance were found between Group A and each of the other three groups, between Group B and Group C, and between Group C and Group D. However, no significant difference was found between Group B and Group D. Because no significant difference was found for Text I between standard cloze (Form D) and the other three cloze forms (Forms A, B, C), any claim about whether examinees taking cloze tests that are based on the text-driven deletion method will perform differently from those taking standard cloze cannot be firmly made in this study. In other words, as the results were different between Text I and Text II, the study did not provide consistent evidence to support the findings of

Alerdson (1979; 1983), Farhady, Jafarpur, and Birjandi (1994), and Farhady and Keramati (1996) that cloze tests with different deletion rates lead to different tests.

On the other hand, one interesting point to note was that large differences in mean score were observed between students taking C-test and those taking the other four cloze forms (see Table 4). For both texts, the mean scores of C-test (68.73 for Text I and 68.97 for Text II) were at least 20 points higher than those of the other four cloze forms. In other words, although all the five test forms belong to the LRR family, C-test appeared to be the easiest for both texts. These large mean differences between the students taking C-test and those taking cloze tests, together with the substantially lower criterion-related validity found above, may thus lend support to Jafarpur's (1995; 1996) claim that C-test seems to measure the ability different from that by cloze test.

In order to examine whether, for each test form, the assumption of test-takers' performance invariance across different texts holds, correlations between test-takers' performance on Text I and on Text II were calculated and are presented in Table 5. A high correlation between the two texts will lend support to the claim that LRR tests using different texts will still rank test takers in similar order. As shown in Table 5, the results indicated that the correlation coefficient for each of the five test forms was positive and significantly different from zero. The correlation coefficients ranged from 0.60 (Form E) to 0.88 (Form B), implying average to moderately high relationship between test-takers' performance on Text I and on Text II. For the purpose of testing whether the relationship between their performance on Text I and on Text II was statistically similar among the five test forms, the five correlation coefficients were transformed to Fisher's Z scores, which are shown in Table 5. Based on the formula in Glasnapp and Poggio (1985), the observed test statistic for each pair of correlation coefficients compared was calculated and compared with the critical Z value of 1.645 with 0.05 significance level. The results indicated that the relationship between test-takers' performance on Text I and on Text II for Form B was statistically different from those for Form A (observed Z value = 2.229), for Form C (observed Z value = 1.735), for Form D (observed Z value = 1.966), and for Form E (observed Z value = 3.192). On the other hand, there was no significant difference in magnitude and direction of the relationship between test-takers' performance on Text I and on Text II among the rest of the four test forms. The findings indicated that test takers taking Form B had the most similar performance on the two texts, meaning that Form B tended to rank test takers' performance in the most similar order across Text I and Text II. In other words, the findings showed that only Form B could provide most evidence for the assumption of test-takers' performance invariance across different texts to hold. However, the results for the rest of the four test forms did not provide strong evidence to prove that cloze tests or C-tests with different texts can be considered parallel. The findings were somewhat in line with those of Alderson (1979) and Zarabi (1988).

Table 5 Correlation (r) and Fisher' s Z Transformation (Z_r) between Test-takers' Performance on Text I and on Text II

Cloze form	No. of students	*r*	Z_r
Form A	49	0.72*	0.908
Form B	49	0.88*	1.376
Form C	41	0.76*	0.996
Form D	54	0.75*	0.973
Form E	44	0.60*	0.693

Note: * $p < 0.05$.

5. Conclusions

Based on the above results, the following conclusions can be made in the order of the four research questions stated earlier: (1) the findings of this study refuted the claim of Farhady and Keramati (1996) that the cloze forms which are based on their text-driven deletion method produce better psychometric properties and are superior to the standard cloze, in which deletion rate is set arbitrarily; (2) contrary to the findings of Klein-Braley (1997), this study found no evidence to show, in terms of criterion-related validity, the superiority of C-test over cloze test; (3) it was still not clear yet about whether different deletion rates lead to difference in test-takers' performance; (4) except the cloze form that is based on noun phrase with the text-driven deletion method, there was no strong evidence to substantiate the claim that various close forms and C-test meet the assumption of test-takers' performance invariance across different texts.

The findings of this study, which refuted the claim of Farhady and Keramati (1996), pointed to one major problem underlying their proposed text-driven deletion method. The problem was that there was no theoretical justification for the formula they proposed to calculate the deletion rate. That is, although taking the linguistic structure of the text into consideration, their formula was formulated purely on an arbitrary basis without regard to other relevant characteristics, such as rhetoric mode, stylistic characteristics, tone, and word choice of the texts. The formula was an oversimplification of the complicated nature of the text and manifested a lack of theoretical rationale for what other relevant component(s) of a text should be included in the denominator of the formula. Hence, the deletion rate obtained using their formula was as arbitrary as the one that was arbitrarily set for standard cloze. Hence, given the fact that this arbitrariness was inherent in both standard cloze and cloze forms that were based on the text-driven deletion method, it was not surprising that conclusion (1) was obtained in this study. The conclusion (1) implies that with a lack of theoretical basis for the formula and a lack of consistent empirical evidence in favor of any of the forms that are based on the text-driven deletion method, it may not be worthwhile to go through all the calculations needed for these cloze forms. Hence, to make things simple, it may be better, at least for the time being, to stick to standard cloze tests when measuring general English proficiency.

Similarly, based on the results of the study, C-test is not necessarily a better choice when one wants to measure overall English proficiency. In fact, the large difference found in mean performance between the various cloze forms and C-test, together with its relatively low levels of agreement with the criterion measures which measure overall English proficiency, seemed to indicate that C-test does not measure the same aspect(s) of English ability as those measured by cloze tests.

However, taking all the findings and conclusions of the study together, one may have to conclude that more definite statement concerning the construct validity of cloze test and C-test can be made only after more well-conducted validation studies are done. A small number of researchers (Sasaki, 2000; Storey, 1997; Yamashita, 2003) have recently conducted validation studies that collected introspective evidence from examinees by asking them to think aloud as they responded to items in the two test procedures. However, their results are quite restricted, as most of their studies were limited in scope, such as including only one validation procedure, one text and a small number of participants from a homogeneous cultural and linguistic background. Therefore, the need to conduct validation studies that include a great variety of texts, a wide range of procedures, and test-takers with different language proficiency levels and diverse cultural backgrounds is certainly warranted. The validation studies may also help to find conclusive results concerning whether both cloze tests and C-tests are robust to text variation or deletion rate variation.

References

Alderson, J. C. (1979). The cloze procedure and proficiency in English s a foreign language. *TESOL Quarterly, 13*, 219-227.

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

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. New York: Heinle & Heinle Publishers.

Bormuth, J. R. (1965). Validities of grammatical and semantic classifications of cloze test scores. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 283-285). Newark, DE: International Reading Associates.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension tests scores. *Journal of Reading, 10*, 291-299.

Carroll, J. B. (1987). Review of Klein-Braley, C. and Raatz, E., editors. 1985. C-Tests in der Praxis. In Fremdsprachen und Hochschule, AKS-Rundbrief 13/14. Bochum: Arbeitskreis Sprachenzentrum (AKS). *Language Testing, 4*, 99-106.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7*, 121-46.

Cohen, A. D., Segal, M., & Weiss, R. (1984). The C-test in Hebrew. *Language Testing, 1*, 70-81.

Crawford, A. (1970). *The cloze procedure as a measure of reading comprehension of elementary level Mexican-American and Anglo-American children*.

Unpublished doctoral dissertation, University of California, Los Angeles, California.

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

Farhady, H., Jafarpur, A., & Birjandi, P. (1994). *Testing language skills from theory to practice*. Tehran: SAMT Publication.

Farhady, H., & Keramati, M. N. (1996). A text-driven method for the deletion procedure in cloze passages. *Language Testing, 13*, 191-207.

Gallant, R. (1965). Use of cloze tests as a measure of readability in the primary grades. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 286-287). Newark, DE: International Reading Associates.

Gamarra, A. G.., & Jonz, J. (1987). Cloze procedure and the sequence of text. In J. E. Readence & R. S. Baldwin (Eds.), *Research in literacy: Merging perspectives* (pp. 17-24). Rochester, NY: National Reading Conference.

Glasnapp, D. R., & Poggio, J. P. (1985). *Essentials of statistical analysis for the behavioral sciences*. Columbus, OH: A Bell & Howell Company.

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. *Language Testing, 2,* 159-185.

Grotjahn, R. (1987). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analysis. In R. Grotjahn, C. Klein-Braley, & D. K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (pp. 219-253). Bochum: Brockmeyer.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. Melbourne: Edward Arnold.

Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. London: Longman.

Henning, G.. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Hinofotis, F. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Jr. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 121-128). Rowley, MA: Newbury House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Jafarpur, A. (1995). Is C-testing superior to cloze? *Language Testing*, *11*, 194-216.

Jafarpur, A. (1996). Native speaker performance validity: In vain or for gain? *System*, *24*, 83-96.

Jonz, J. (1989). Textual cohesion and second-language comprehension. *Language Learning*, *37*, 409-38.

Klein-Braley, C. (1981). *Empirical investigations of cloze tests*. Unpublished doctoral dissertation, University of Duisburg, Duisburg, Germany.

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

Lee, P. T. (1988). *A reader for writers*. Taipei: Bookman Books Ltd.

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

Ohnmacht, F. W., Weaver, W. W., & Kohler, E. T. (1970). Cloze and closure: A factorial study. *The Journal of Psychology*, *74*, 205-17.

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

Oller, J. W. Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Raatz, U., & Klein-Braley, C. (1981). The C-Test: A modification of the cloze procedure. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 113-148). Colchester: University of Essex.

Ruddell, R. B. (1964). A study of the cloze comprehension technique in relation to structurally controlled reading material. *Improvement of Reading through*

Classroom Practice, *9*, 298-303.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing*, *17,* 85-114.

Spolsky, B., Bengt, S. M., Sako, E. W., & Aterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. In J. A. Upshur & J. Fata (Eds.), *Problems in foreign language testing, language learning special issue* (pp. 79-103). Ann Arbor, MI: Language Learning Research Club, University of Michigan.

Steig, W. (1969). *Sylvester and the magic pebble*. NY: Simon & Schuster Children's Publishing Division.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing*, *14*, 214-31.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.

Weaver, W. W., & Kingston, A. J. (1963). A factor analysis of cloze procedure and other measures of reading and language ability. *The Journal of Communication*, *13*, 252-61.

Weir, F. J. (1988). *Communicative language testing*. Exeter: University of Exeter.

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

Zarrabi, A. (1988). *Deletion rate and test difficulty in cloze*. Unpublished master's thesis. Allameh Tabatabai University, Tehran, Tehran, Iran.

Appendix A: Sylvester and the Magic Pebble (Extraction)

Sylvester Duncan lived with his mother and father at Acorn Road in Oatsdale. One of his hobbies was collecting pebbles of unusual shape and color. On a rainy Saturday during vacation he found a quite extraordinary one. It was flaming red, shiny, and perfectly round, like a marble. As he was studying his remarkable pebble, he began to shiver, probably from excitement, and the rain felt cold on his back. "I wish it would stop raining," he said.

To his great surprise the rain stopped. It didn't stop gradually as rains usually do. It ceased. The drops vanished on the way down, the clouds disappeared, everything was dry, and the sun was shining as if rain had never existed.

In all his young life Sylvester had never had a wish gratified so quickly. It struck him that magic must be at work, and he guessed that the magic must be in the remarkable-looking red pebble. (Where indeed it was.) To make a test, he put the pebble on the ground and said, "I wish it would rain again." Nothing happened. But when he said the same thing holding the pebble in his hoof, the sky turned black, there was lightning and a clap of thunder, and the rain came shooting down.

He wished the sunshine back in the sky, and he wished a wart on his left hind fetlock would disappear, and it did, and he started home, eager to amaze his father and mother with his magic pebble. He could hardly wait to see their faces. Maybe they wouldn't even believe him at first.

As he was crossing Strawberry Hill, thinking of some of the many, many things he could wish for, he was startled to see a mean, hungry lion looking right at him from behind some tall grass. He was frightened. If he hadn't been so frightened, he could have made the lion disappear, or he could have wished himself safe at home with his father and mother.

He could have wished the lion would turn into a butterfly or a daisy or a gnat. He could have wished many things, but he panicked and couldn't think carefully.

"I wish I were a rock," he said, and he became a rock.

The lion came bounding over, sniffed the rock a hundred times, walked around and around it, and went away confused, perplexed, puzzled, and bewildered. "I saw that little donkey as clear as day. Maybe I'm going crazy," he muttered.

And there was Sylvester, a rock on Strawberry Hill, with the magic pebble lying right beside him on the ground, and he was unable to pick it up. “Oh, how I wish I were myself again,” he thought, but nothing happened. He had to be touching the pebble to make the magic work, but there was nothing he could do about it.

Appendix B: Do You Want to be Wise? Rich? Famous?

“God says: Take what you want and pay for it!” When I first heard this proverb from Spain it frightened me; I used to dream of an Angel with a flaming sword. But as I thought more about it, I realized that the Angel held not a sword but a balance.

In one side, you put what you would like to be. Do you want to be famous? Very well, says the Angel, then spend every waking hour in the pursuit of fame. It will show up on the other side of the balance in time spent and sacrifices made. Is it riches you want? Think about money every day, study it, give your life to it, and the balance will be weighted with gold---but at the cost of other things.

Maybe you want to be wise. The Angel will weigh out a high payment for that, too; it will include a good life, a pursuit of knowledge, and an uncompromising love of truth.

Everything has its price. We are familiar with this idea in our daily lives. We go to the self-service store. In our wire cart we put a can of tomatoes, a bit of cheese, bread, hamburger and spaghetti. On the way out the clerk adds up our bill, puts our purchases in a paper bag, and we carry home our dinner---after we have paid for it.

So with the balance of our lives: on one side, our heart’s desire; on the other side of the scales, the reckoning. When the scales are even, you may take out what you have bought. Sometimes the price seems high. But remember, you must pay for the character and quality of your goal as well as for the achievement of it. The law is simple and it is just; you may have what you want---but you must pay. Nothing is free.

Language Reduced Redundancy Tests: A Reexamination of Cloze Test and C-test

Wen-Ying Lin *　Hsiao-Ching Yuan**
Ho-Ping Feng***

The purposes of this study are: (1) to investigate whether cloze forms with text-driven deletion method, proposed by Farhady and Keramati (1996), will produce better psychometric properties than the standard cloze form; (2) to compare the psychometric properties of cloze test and C-test, both of which belong to the family of language reduced redundancy test; (3) to examine whether different deletion rates lead to difference in test-takers' performance; and (4) to test whether the assumption of test-takers' performance invariance across different texts hold for both cloze test and C-test. Based on two authentic texts with different rhetoric modes, three cloze forms with text-driven deletion method, along with one standard cloze and one form of C-test, were constructed and randomly administered to 237 student subjects at one private university in northern Taiwan. Furthermore, each subject was required to take three subtests (from a sample TOEFL test) as criterion measures for empirical validity. The results of the study indicated that neither the three cloze forms nor the C-test was substantially superior to the standard cloze in terms of reliability and validity. In addition, the findings of the study were inconclusive with regard to whether different deletion rates result in different test-takers' performance. Finally, no strong evidence was found to substantiate the claim that both cloze test and C-test meet the assumption of test-takers' performance invariance across different texts.

Keywords: language reduced redundancy test, cloze test, deletion rate

* Wen-Ying Lin, Associate Professor, Department of English Instruction, Taipei Municipal University of Education.

**Hsiao-Ching Yuan, Lecturer, Language Center, Ming Chuan University.

***Ho-Ping Feng, Associate Professor, Department of English, National Taiwan Normal University.