Modifying C-Test for Practical Purposes

Kenji ISHIHARA, Elizabeth HISER, and Tae OKADA

Keywords: versions, equivalency, readability, validity, reliability

Abstract: This paper reports on a project of writing a C-Test which would allow all students at Doshisha University to fully demonstrate their English-language ability. Practical reasons in administering the new C-Test necessitated modifications of the C-Test procedure set forth by Raatz and Klein-Braley. The resulting "fixed-ratio" C-Test had three versions with the last half of every fifth or sixth word to be supplied by the test-taker. Evidence on inter-version equivalency is presented on the basis of pilot test results with some 200 students.

Introduction

In 1993, two versions of C-Test were developed by the Writing Research Group (WRG) of the Kansai Chapter of the Japan Association of College English Teachers (JACET) (WRG, 1995). Since then, both versions of the C-Test have been used in numerous EFL research projects by the WRG's members proving repeatedly to be a practical and efficient means of measuring the general proficiency levels of college students in the Kansai area. In 1998, the two C-Tests underwent a major revision and continued to be used (WRG, 1995, 1997, 1998, 1999a, 1999b; Ishihara 2000). The

Doshisha Studies in Language and Culture, *5*(4), 2003: 539 – 568.
Doshisha Society for the Study of Language and Culture,

scores obtained over the years from the participating students indicated that the textual difficulty levels of the two C-Tests were adequate for a wide range of English-language proficiency levels of Japanese college students.

At Doshisha University, however, every time the C-Test was administered, there were a few students who obtained perfect score on both versions of the C-Test. This meant that the C-Test did not give these students an opportunity to demonstrate their entire English-language ability. Thus, it became necessary to revise the C-Test again in such a way that it would measure every single participant's proficiency to the full. Since both versions of the C-Test were comprised of four short passages of slightly different textual difficulty levels (See Appendix 1), two directions were envisioned for developing new C-Tests. One was to select entirely new passages and write completely new C-Tests, and the other was to retain one or more of the four existing passages and add one or more new passages of a slightly greater complexity to form modified versions of the C-Tests.

The present paper describes, step by step, the course of action taken for the first of these two alternatives, namely writing an entirely new C-Test.[1]

Backgrounds

C-Test was first proposed in 1981 (See Dornyei & Katona, 1992) by Raatz and Klein-Braley as a "general language proficiency" test (Carroll, 1987). It uses several (typically from four to six) "short [presumably approximately 100-words each in length], carefully selected, preferably authentic texts. . . [with] . . . the second half of every other word [deleted] beginning from word two in sentence two" (Klein-Braley, 1997: p. 64). (See also Raatz and Klein-Braley, 1996 or website.) The entire test should have at least 100 deletions, "around 90%" of which must be restored correctly by "a control group of adult educated native speakers or teachers of the language. . . . Only entirely correct restorations are counted as

correct" (Klein-Braley, 1997: p. 64).[2]

C-Test is one of the four different types of cloze tests, the other three being the fixed-ratio cloze, the rational cloze, and the multiple-choice cloze. Of these four, the most difficult is said to be the fixed-ratio cloze, while the easiest was shown to be the multiple-choice cloze; C-Test is considered to be the second easiest (Chapelle & Abraham, 1990). C-Test procedure is reported to be useful for languages other than English such as German, French, Spanish, Hebrew (Carrol, 1987; Cohen, Segal & Bar-Siman-Tov, 1984), and, with necessary modifications, Japanese (Hata, 1990).[3]

In a number of studies, C-Test scores have been demonstrated to correlate highly with other language tests. Nigishi (1987) compares his C-Test results with Japanese university students with the English Language Battery (ELBA), Part II, concluding that C-Test "appears to be both a reliable and valid measure of general language proficiency" (p. 24). Chapelle & Abraham (1990) correlate their results with vocabulary, reading, writing and listening tests. Dornyei and Katona (1992) compare their C-Test data with vocabulary, grammar, and listening comprehension test results as well as TOEIC scores. Hastings (website) presents correlations between C-Test (constructed at the University of Wisconsin-Milwaukee referred to above) scores and the TOEFL scores. Mochizuki (1994) compares her C-test data both with the Second Grade Test of the Society of Testing English Proficiency (STEP) and Comprehensive English Language Test for Learners of English (CELT) in addition to her own listening and dictation test scores. Ikeguchi (1998) correlates her C-Test results with the STEP results. Writing Research Group of JACET Kansai Chapter (1995, 1997, 1998, 1999a, 1999b) repeatedly obtained data on correlations between C-Test and writing fluency scores among a wide-range of Japanese college students. Ishihara, Okada and Matsui (1999, 2000) also showed close correlations between their vocabulary survey data and C-Test scores.

Carroll (1987) points out that "the main problem [with C-Test

construction] seems to be the selection of passages . . . usually . . .[conducted] on the basis of intuitive judgments about difficulty and content." In relation to the difficulty levels of test passages, Tsuchiya (1998) finds C-Tests useful "as a measurement tool of readability" (p. 200). Kamimoto (1992) presents evidence demonstrating that both familiarity of topics and textual readability influence the C-Test scores, adding that "to get a good random sample of [test] words . . . there is a need to modify [the C-Test] deletion procedure" (p. 76).

Kamimoto (1993) "tailors" his original C-Test by retaining the test items with item facility between .29 and .71 and eliminating those with lower item discrimination than .20; his "tailored C-Test turned out to be more reliable than the original version" (p. 52). Kakkota (1988) "rationalizes" the C-Test deletion rules both in inter-item distance (IID) and number of letters deleted, which varies from zero to half the word (minus one or two letters when the item word is longer than six letters (p. 116), concluding that "the optimal average IID appears to be five words" (p. 118). Mochizuki (1994) compares scores obtained for four types of approximately 400-word passages, i.e., expository, argumentative, descriptive and narrative, with 120 deletions each. She recommends that the C-Test be made out of a single narrative text of around 400 words in length. Jafarpur (1995) experiments with "20 C-tests . . . constructed from the same text, each distinct in deletion start and/or deletion ratio," (p. 197) arriving at the conclusion that "by utilizing different ratio and deletion starts, the [C-Test] procedure produces tests that are widely apart" (p. 205).

Writing C-Test passages

The first stage in writing a C-Test is to choose appropriate passages. A number of passages of various length on a variety of topics were first collected as candidates. In order to make the new C-Test slightly higher in

proficiency level than the earlier versions, i. e. those published in Writing Research Group (1999a: pp. 81-85), the passages of the new C-Test had to be more complex in content and language, which, in the end, called for longer passages than the earlier ones. The original passages had 55-59 words per passage, while the new passages were approximately 125 words in length on the average (Appendix 1). For that reason, the new C-Test had to be limited to three topics or passages for practical purposes.

After the three passages were selected, they were revised for unity, brevity, and clarity (See Appendix 2 for the test passages). They were arranged on the basis of the readability scores as well as intuition with the easiest first, the most difficult second, and the middle one last. From many years of classroom experience, this arrangement of the passages was considered best for guaranteeing the morale and steady effort on the part of the students.[4]

Once the passages were established, it was realized that deleting the last half of every other word would give far too many blanks to be filled by the test-takers. Based on past experience with C-Tests, the testing time was considered best for practical purposes not to exceed 15-20 minutes. As a consequence, a reasonable number of test items was to be no more than fifty. For the newly selected passages, the ratio of deletion had to be reduced to every fifth or sixth word instead of every second. Thus, the new C-Test was to be a "fixed-ratio" C-Test with the last half of every fifth or sixth word to be supplied by the test-taker.

The resulting test obviously did not follow the recommendations set forth by Raaz and Klein-Braley (1996) for C-Tests. The new C-Test had three passages that were slightly longer instead of "a number of" very short passages. Although Raaz and Klein-Braley did not specify the number of passages nor the length of each passage, they indirectly suggested 5 or 6 passages with 20-25 blanks each. A passage with 20-25 blanks meant 40-50 words in length (because of "the rule of two") plus one sentence at the

beginning and perhaps another at the end, so that the total number of words per passage probably would not exceed 100 words. These new passages were not arranged from the easiest progressively to the most difficult as proposed by Raaz and Klein-Braley. Regarding the total number of test-items, Raaz and Klein-Braley explicitly recommended "at least 100 items," while ours had fifty. For a Cloze test, on the other hand, Taylor (1953) proposed the minimum number of blanks to be 50 "in order to ensure adequate sampling" (Raaz & Klein-Braley, 1996: p. 3). Unlike Raaz and Klein-Braley's "rule of two," the test developed in the present project deleted the second half of every fifth or sixth word, selecting the first test item somewhere between the second and the tenth word in the second sentence of the passages.

These decisions, in turn, made it possible to develop three different versions of C-Test by deleting different words from the same passages. Thus, in the second sentence of the first passage, for instance, "It is known that the suspect is a man in his early thirties," the first test item of the three different versions was to be the fourth word "that," the sixth word "suspect," and the ninth word "man" respectively, and subsequent test items were every fifth or sixth word from there.

Research issue

Where there are three versions that differ only in the deleted items, the question arises as to the equivalency of the three versions. As pointed out by Carroll (1987), "the selection of passages" is a major factor in C-Test construction. The score average on any C-Test is known to differ not only according to the degree of complexity or readability of test passages (Tsuchiya 1998), but also to "familiarity of topics" (Kamimoto 1993) and textual types (Mochizuki 1994). When the passages are identical, however, test scores are not influenced by the text; thus if the test scores differ

significantly among the versions, the difference would only be due to the test items. By testing students with various versions of C-Test with identical passages, an opportunity is provided to clarify the extent to which the test items–and not the text–play on the test scores. If the three versions prove to be close equivalents, it would mean that C-Test score is more dependent on text than on items. If, on the other hand, the test scores differ from version to version, it could be assumed that the C-Test score is more item-dependent than text-dependent. Put differently, inter-version equivalency, or lack thereof, would determine the extent to which the test items influence the total scores.

Validity of the modified C-Test

In order to examine face or content validity of the new C-Test, the three versions were presented to a total of thirty native speakers of American English to verify whether they could complete the passages with "an acceptable level of accuracy: around 90% correct on average" (Klein-Braley, 1997: p. 64). All the thirty informants, each of whom took one of the versions, scored higher than 92% with average scores as shown in Table 1.

In order to further test face or content validity, the test items were analyzed in terms of parts of speech or grammatical functions. As shown in Appendix 3, the test items of each version presented a wide variety of parts of speech and grammatical functions. Therefore, face or content validity seemed apparent.

For criterion validity, the Michigan English Placement Test (MEPT) was administered in two of the classes as well as the modified C-Test. The comparison of the results proved that "reading ability . . . is the best predictor of abilities measured by the C-Test, although all four of the MEPT section scores [i.e. reading, vocabulary, structure, and listening] do sustain

Table 1: Mean scores of the native-speaking informants

Versions	Number of informants	Passage 1	Passage 2	Passage 3	Total score
Full score	—	30	30	40	100
Version 1	11	30.0	29.0	38.1	97.1
Version 2	11	29.9	28.0	38.2	96.1
Version 3	8	29.4	27.9	38.5	95.8
Average in %	—	99.3	94.3	95.8	96.3

significant relationships with all versions of the C-Test" (Hiser, Ishihara & Okada (forthcoming)).[5] It is to be noted that "the C-Test is capable of . . . measuring, to a certain extent, listening ability" as well (ibid.).

Pilot-test participants

The three versions of the new modified C-Test were administered in nine second-year English-language classes where the students were registered to fulfill a language requirement. These were non-tracked or un-streamed students, so the range of proficiency in English was wide. The participants took one version at a time: the first version, for instance, as part of the pre-tests at the beginning of the academic year, the second version as part of the first-semester final or the second-semester pre-test, and the third version as part of the final exam at the end of the academic year. In view of the later inter-version correlation analysis, the three versions were administered in varied orders for maximally equalizing the resulting scores. Five of the classes took all three versions in different orders, and the other four took two of the three in different combinations, e. g., one class took Versions 1

Table 2: Number and faculty affiliation of the pilot-test participants

Faculty	Theology	Letters	Law	Economics	Commerce	Engineering	Total
Version 1	2	32	48	49	47	56	234
Version 2	1	21	35	43	28	58	186
Version 3	1	27	60	38	45	50	221

and 2, a second class, Versions 2 and 3, etc. The total number of participating students and their faculty affiliation was as shown in Table 2.

Scoring procedures

Scoring the participants' responses was conducted manually with close cooperation among the three authors of this paper. If the deleted half of the test words was correctly restored, it was assessed as two points. Since there were 50 test items in each version, the full score was 100 points. If a form given by a participant was wrong in meaning in the given context, it was given no point: "final" instead of "first," for instance, or "world" instead of "word." On the other hand, if responses differed from the expected forms in one of the following ways, the three scorers consulted each other before assessing them for 1 point:

1) Forms with a spelling error correctable on a standard word processor;
2) Otherwise acceptable forms with more or fewer letters than allowed for the blank than required;
3) Correct forms except for the word ending such as the plural and past-tense suffixes;
4) Words that are almost but not quite appropriate in meaning in the given context.

In brief, the overall approach was that if the students indicated that they understood what the word should be but made minor mistakes in form or spelling, they were given half credit.

As a way of measuring the severity of spelling errors, the relevant forms were tested on a widely-used word processing computer program, that is, if they were readily correctable on the computer program, the error was considered minor enough to deserve 1 point: "marcket" for "market," or "lenkth" for "length," for example. Some of the unexpected responses

could be acceptable in the given context, thus both “gun” and “guns” would do in “he is still carrying the g____,” and both “applied” and “applicable” in “An important lesson that was first learned about advertising on radio was appli____,” but the “rule” of the C-Test required only 1 or 2 letters in the first example, and 5 or 6 letters in the second. For that reason, both “guns” and “applied” were assessed for 1 point because the former has one more letter and the latter, three or four fewer letters than the C-Test rule dictated.

In actual practice, however, not everything could be done so mechanically. For instance, in the sentence “Advertisers could now picture the product,” some participants responded with “advertising,” which was not the expected form but had to be accepted as correct. Similarly, both “form” and “focus” were considered good in “the fo____ of the ad was at least as important as the content.”[6] Another example where an unexpected form was accepted as good was the item “isn’t” for “It i____ difficult to imagine ....” While “is” was the expected form, a few pilot test participants and one of the native-speaking informants filled this blank with “isn’t.” Taking the native speaker’s response into consideration, an exception was made for this test item to accept “isn’t” as correct although it exceeded the number of letters allowed for this blank. For the last word in that sentence, “content,” some participants gave the alternative forms, “context” or “concept,” both of which were considered awkward but not totally unacceptable, so that both were given 1 point.

In sum, scoring was not as simply mechanical as one might have expected. It could, of course, be made simple by categorically rejecting virtually every unexpected form, known as the “exact method” (Raatz and Klein-Braley, 1996), but this was felt to reduce accuracy of assessment.

Pilot test results

General statistics regarding the pilot test results are shown in Table 3.

The mean score was highest for Version 1 at 56.60 and lowest for Version 2 at 47.82, while the standard deviation was also highest for Version 1 at 15.90 and lowest for Version 3 at 14.59. In view of the fact that the mean scores were around 50% as proposed by Raaz and Klein-Braley (1996 or website),[7] the difficulty levels of the test passages seem adequate for the target group of students, although the range from the maximum to minimum scores is wider than between 20 and 80, the range suggested by Raaz and Klein-Braley (1996 or website). Besides, since no test participant obtained a perfect score, one of the purposes of the project proved to be successful, namely to provide every student an opportunity to demonstrate their English proficiency to the full.[8] In kurtosis, the largest deviation from zero, i. e. -0.46, was seen for Version 2 probably due, at least partly, to the smaller number of participants than for the other two versions (See the "Discussion" section below for details). The reliability score, Cronbach's Alpha, was uniformly above 0.85, close to Raaz and Klein-Braley's aim of 0.90 (Raaz & Klein-Braley, 1996 or website).

Table 3: General statistics of the pilot test

Versions of C-Test	Version 1	Version 2	Version 3
Number of participants	234	186	221
Maximum	90	84	98
Minimum	7	7	11
Mean	56.60	47.82	55.55
Standard Deviation	15.90	15.83	14.59
Kurtosis	0.298	-0.464	0.071
Cronbach's Alpha	0.8891	0.8719	0.8641

Rescoring the pilot test results

Among the fifty test items in each version were several relatively easy ones as well as difficult ones. When an item was easy, top scorers as well as low scorers filled the blanks successfully, and therefore, such items did not differentiate the top scorers from low scorers. On the other hand, if an

item was difficult even for top scorers to respond with success, this again failed to show appropriate difference between the high- and low-scoring groups. In order to clarify which of the test items discriminated effectively the top from the low scorers, the percentage of successful responses or item facility (IF) for the top third and the low third of the participants was computed, and the difference between the top group's IF and low group's IF or item discrimination (ID) was obtained for each of the test items. Subsequently, those test items with lower ID than 0.3 were deleted as less efficient discriminators, and the total test results were rescored. Table 4 summarizes the general statistics of the rescored data. It is to be noted that the correlations between the original total scores and the rescored totals were 0.990 for Version 1, 0.992 for Version 2, and 0.977 for Version 3 (Table 4), again demonstrating the high reliability of the original scores.

After rescoring, the correlations between the total and passage scores were examined (lower half of Table 5). They were all higher than 0.83, and especially high for Passage 3 which were above 0.9. They differed only minimally after rescoring (Table 6). This meant that all three passages in all three versions contributed highly to the total scores.

The inter-version correlation figures were invariably higher than 0.6 and slightly higher for rescored totals between Versions 1 and 3, and Versions 2

Table 4: General statistics of the rescored pilot test results

Versions of C-test	Version 1	Version 2	Version 3
Numberof participants	234	186	221
Number of test items	37	40	31
Perfect score	74	80	62
Maximum	72	74	62
Minimum	5	6	6
Mean	44.75	39.74	36.82
Mean in percentage[9]	60.47	49.68	59.39
Standard Deviation	13.96	14.56	11.76
Kurtosis	-0.121	-0.592	-0.573
Cronbach's Alpha	0.8775	0.8694	0.8488
Correlation with original total	0.990	0.992	0.977

and 3 (Table 7). From these correlation figures, the three versions might well be assumed to be reasonably equivalent.

Table 5: Correlations between the total score and the passage scores before and after rescoring

Correlations between the total and passage scores before rescoring			
Versions	**Version 1**	**Version 2**	**Version 3**
Passage 1	.859 (0.882)	.837 (0.850)	.810 (0.824)
Passage 2	.849 (0.867)	831 (0.831)	.840 (0.857)
Passage 3	.910 (0.918)	937 (0.936)	.905 (0.918)
Correlations between the total and passage scores after rescoring			
Passage 1	.858 (0.882)	.840 (0.848)	.807 (0.811)
Passage 2	.855 (0.860)	. 830 (0.831)	.837 (0.835)
Passage 3	.914 (0.909)	. 948 (0.947)	.898 (0.906)

SPSS figures followed by Excel figures in parentheses; all correlation figures are significant at P≦.01

Table 6: Difference in passage-total correlations before and after rescoring

Versions	**Version 1**	**Version 2**	**Version 3**
Passage 1	0	-0.002	-0.013
Passage 2	-0.007	0	-0.022
Passage 3	-0.009	+0.011	-0.012

Table 7: Correlation among total scores of Versions 1, 2, and 3

Correlation among total scores before rescoring		
	Total score of Version 1	**Total score of Version 2**
Total score of Version 2	.610 (0.630)	——
Total score of Version 3	.677 (0.681)	.662 (0.680)
Correlation among total scores after rescoring		
	Rescored total of Version 1	**Rescored total of Version2**
Rescored total of Version 2	.609 (0.626)	——
Rescored total of Version 3	.687 (0.698)	.682 (0.695)

SPSS figures followed by Excel figures in parentheses; all correlation figures are significant at P≦.01

Discussion on the pilot test results

The first question on pilot test results concerned the average total scores of the three versions, especially the difference between the mean score of Version 2 and those of the other two versions. Potential causes for the lower mean score in Version 2 could have been any of the following:

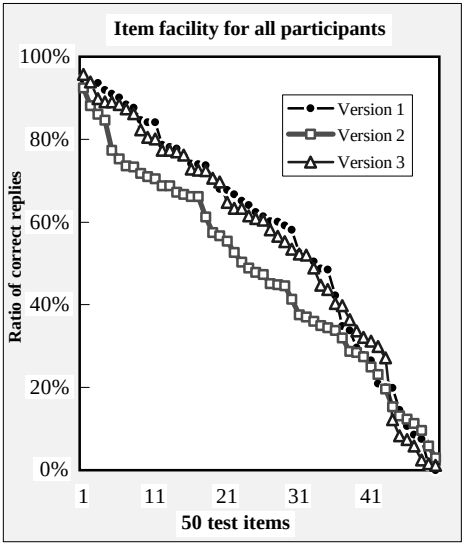
(1) The test items were more difficult in Version 2.

(2) High-scoring participants in the other versions did not participate in Version 2.

(3) Some degree of memorization or practice effect involved in the scores of Versions 1 and 3 was missing in Version 2.

In order to clarify whether or not the test items were more difficult in Version 2, IF or the percentage of correct responses on the three versions were compared. (See Figure 1, in which the IF figuers are arranged from the highest on the left to the lowest on the right.) As visible in Figure 1, a large majority or 38 of the 50 items in Version 2 were lower in IF than for Versions 1 and 3. The characteristics of the 38 seemingly more difficult items of Version 2 were compared with those of the 12 others as well as the items in the other versions. No specific cause was discovered on the surface to make the 38 items more difficult, mainly because they were as varied and wide-ranging as in the other two versions in grammatical and/or lexical

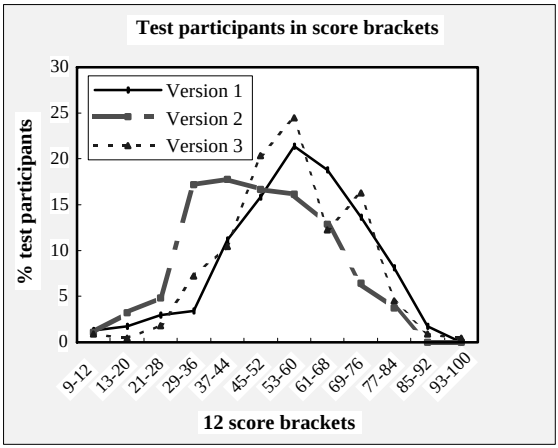
Figure 1: Item facility for the 50 test items of Versions 1 through 3



nature.

For further investigation of the cause for the lower mean score in Version 2, the profile of the entire group of participants was examined. For this purpose, the test-takers were categorized into 12 score brackets from zero to 100 points (the horizontal axis of Figure 2). Thus for Version 1, for instance, the largest number or 21.4% of the participants were in the 7th category of 53-60 points. Evidently in Version 2 there were fewer participants in the high score brackets above 52 points, and a larger number of participants in the brackets below 45 points than for Versions 1 and 3. Therefore, for the purpose of comparison, the participants were divided into two groups, namely, those who scored higher than 52 points and lower than 45 in all three versions, and comparison was made between high-scoring and low-scoring participants in the three versions (Table 8). The mean scores were 65.20, 63.59, and 64.25 for Versions 1, 2, and 3 respectively for those who scored 52 points or more; the difference in the mean scores between Versions 1 and 2 on the one hand, and between Versions 2 and 3 on the other were 1.61 and 0.66 respectively. For those who scored 42 or

Figure 2: Pilot test participants in score brackets



lower, the mean scores were 36.16, 33.45, and 36.44 respectively for the three versions. The differences in the mean scores were 2.71 between Versions 1 and 2, and 2.99 between Versions 2 and 3, both higher than the differences of the mean scores for those above 52 points (Table 8).

Of the 157 participants in Version 1 and 138 in Version 3 who scored 52 points or higher (Table 8), 85 or 54.1 % of the former and 73 or 52.9% of the latter participated in Version 2, in which not all scored higher than 52 points (Table 9). This meant that the other 45.9% and 47.1% of high scorers in Versions 1 and 3 respectively did not participate in Version 2, contributing to lowering the mean of Version 2. On the other hand, of the

Table 8: Statistics for high (52 or above) and (low 45 or below) scorers in Versions 1 through 3

Groups	Versions	Version 1	Version 2	Version 3
High scoring Group (52 points or above)	N	157	74	138
	Mean	65.20	63.59	64.25
	Maximum	90	84	98
	Minimum	52	52	52
	Difference in mean scores	Between Versions 1 and 2: 1.61 Between Versions 2 and 3: 0.66		
Low scoring Group (45 points or lower)	N	57	83	52
	Mean	36.16	33.45	36.44
	Maximum	45	45	45
	Minimum	7	7	11
	Difference in mean scores	Between Versions 1 and 2: 2.71 Between Versions 2 and 3: 2.99		

Table 9: Participants with a total score of 52 or above on Versions 1 and 3

	Version 1	Version 2	Version 3
N	**157**	85	—
Mean	**65.20**	54.76	—
Maximum	**90**	84	—
Minimum	**52**	16	—
Difference in mean scores		10.44	
N	—	73	**138**
Mean	—	54.75	**63.59**
Maximum	—	84	**84**
Minimum	—	16	**52**
Difference in mean scores		8.84	

83 participants scoring 45 points or lower in Version 2 (Table 10), 56 students or 67.5% took Version 1 (not necessarily scoring less than 45 in Version 1) whose mean score was 44.41 points, while 52 students or 62.7% participated in Version 3 (not necessarily scoring less than 45 in Version 3) whose mean score was 45.25 points (Table 10). Both these mean scores were higher than 33.45 points which was the mean score of the 83 low-scoring participants in Version 2 (Tables 8 and 10). Here again, the mean for Version 2 turned out to be lower than those for the other two versions.

In order to make score comparisons more exact, the lists of participants were matched between Versions 1 and 2 on the one hand, and Versions 2 and 3 on the other, by deleting those who took one but not the other. The same process of deletion and matching was used both for the high- and low-scoring participants. The differences in the mean scores obtained in this manner widened between Version 2 and the other two Versions in all cases but one, i. e. between Versions 1 and 2 for high scorers. In this single case, because the mean scores of the 85 participants in both Versions 1 and 2 was 61.71 in Version 1 (Table 11) instead of 65.20 which was for all the 157 participants (Table 9), the difference between the two versions narrowed from 10.44 to 6.95 (Tables 9 and 11). But in the other three cases, the difference between the mean scores of Version 2 and the other two versions widened slightly. Specifically, between Versions 2 and 3, for instance, the high-scoring 73 participants obtained an average of 54.75 and 65.30 respectively (Table 11), while the average for all the 138 Version 3

Table 10: Participants with a total score of 45 or below on Version2

	Version 1	Version 2	Version 3
N	56	**83**	52
Mean	44.41	**33.45**	45.25
Maximum	75	**45**	74
Minimum	7	**7**	20
Difference in mean scores	Between Versions 1 and 2: 10.96 Between Versions 2 and 3: 11.80		

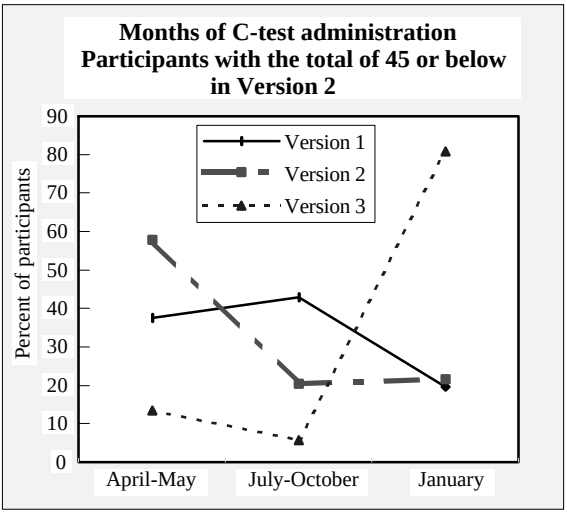
Table 11: Mean scores for identical sets of high- and low-scoring participants

		Version 1	Version 2	Version 3
High scorers	N	85	85	—
	Mean	61.71	54.76	—
	Maximum	75	84	—
	Minimum	52	16	—
	N	—	73	73
	Mean	—	54.75	65.30
	Maximum	—	84	98
	Minimum	—	16	52
	Difference in mean scores	Between Versions 1 and 2: 6.95		
		Between Versions 2 and 3: 10.55		
Low scorers	N	56	56	—
	Mean	44.41	32.45	—
	Maximum	75	44	—
	Minimum	7	7	—
	N	—	52	52
	Mean	—	32.87	45.25
	Maximum	—	43	74
	Minimum	—	7	20
	Difference in mean scores	Between Versions 1 and 2: 11.96		
		Between Versions 2 and 3: 12.38		

participants was 63.59 (Table 9); the difference, therefore, was wider for the same 73 participants in these two versions, i. e. 10.55 (Table 11) than for all 138 participants, i. e., 8.84 (Table 9). This meant that when the same sets of student scores were compared, their averages were generally even higher for Versions 1 and 3 than for Version 2.

For the low scorers also, the exact same 56 participants in Versions 1 and 2 obtained mean scores of 44.41 and 32.45 respectively with the difference of 11.96 (lower half of Table 11), which was larger than 10.96 when all 83 low-scoring participants' mean score was compared with that of all 56 low-scoring participants in Version 1 (Table 10). Similarly, for the identical list of 52 students who took both Versions 2 and 3, the mean scores were 32.87 and 45.25 respectively, the difference being 12.38 (lower half of Table 11); again the difference widened from when the mean score for all 83 Version 2 participants (33.45) was taken for comparison, that difference being 11.80 points (Table 10). This confirms the fact that the same students obtained higher scores in Versions 1 and 3 than in Version 2. In sum, in every way

Figure 3: Percentage of low scorers for three periods of C-Test administration



the scores were examined in relation to the participants, the mean score was lower for Version 2 than for the other two versions.

The question still remained to test the third and last potential cause for the lower mean score for Version 2, namely the possibility of memorization or practice effect resulting from different months in the course of the academic year when the three versions were administered. Since the three versions were administered in different orders to different classes at three different times, i.e. at the beginning (April-May), in the middle (either July or October), and at the end (January) of the academic year, the number of participants in the three versions was calculated separately for the three time periods of the year. This was based on the hypothesis that the version presented later in the year might involve some degree of memorization or practice effect, especially because the three versions were based on identical passages with only different words deleted. Therefore, the number and percentage of participants in the three versions were computed for the three time periods of the year. The result as shown in Figure 3 indicated that

Table 12: Rescored mean for April-May, July-October, and January

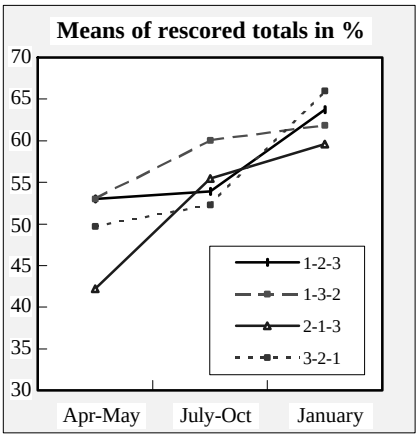
Order of Versions	N	April-May Rescored Mean %	July-October Rescored Mean %	January Rescored Mean %
1-2-3	15	52.99	53.92	63.76
1-3-2	13	52.99	60.05	61.83
2-1-3	37	42.20	55.44	59.59
3-2-1	17	49.72	52.28	65.98

nearly 60% of the participants in Version 2 took it at the beginning of the academic year, while more students took Version 1 in July or October. It is to be noted that about 80% of Version 3 participants took it at the end of the academic year. Thus, the score averages for Versions 1 and 3 could reasonably be assumed to reflect at least some degree of practice effect or memorization.

In an attempt at determining the extent of the score enhancement toward the end of the academic year, the mean scores were calculated for the three time periods of the C-Test administration, namely April-May, July-October, and January. First, the 82 participants who took the C-Test at all three times during the year were divided into four groups according to the order in which the three versions were taken (Table 12). The first group of fifteen students took Version 1 in the April-May period, Version 2 in July-October, and Version 3 in January; the second group of thirteen took Version 1 first, Version 3 second, and Version 2 at the end of the academic year. The third group of thirty-seven participants were administered the C-Test in the order of Version 2 first, Version 1 second, and Version 3 last, while the fourth group of seventeen took the C-Test in the order of Version 3, Version 2, and Version 1. The rescored means for each version are shown in Table 12.

For all four groups of participants, regardless of the order of versions, the mean scores rose invariably toward the end of the academic year. Although part of the score improvement must be due to enhancement of proficiency levels in the course of the academic year, no good method could be found to

Figure 4: Mean scores of the rescored totals at three time periods



isolate that factor from the influence of repeated exposure to the same test passages. Thus, it could only be assumed that at least part of the mean score improvement might have derived from practice effect or memorization. Interestingly, the mean score for Version 2 in April-May was noticeably lower than for the other two versions, reflecting or accounting for, at least in part, the relatively low mean score in Version 2 for the entire sample of 186 participants (Table 3), as well as the lower IF figures for 38 items in Version 2 (Figure 1).

In brief, the relatively low mean score for Version 2 might well be due, at least in part, to the smaller sample (i. e., high-scoring students not participating in Version 2) and lack of practice effect or memorization (i. e., more students took Version 2 earlier in the year), although the possibility could not be totally excluded at this point that the test items in Version 2 were more difficult than those in the other two versions (Figure 1). With that reserve, the three versions could be considered reasonably close equivalents with relatively little effect from the test items. This aspect of Version 2, however, might deserve further investigation.

Summary and conclusion

The modified C-Test was made with three passages of 112-143 words respectively, in which the last half of every fifth or sixth word was deleted for the test takers to supply. By selecting different test words, three different versions were developed with the same three passages. Both content and criterion validity proved adequate for the three versions. Some 200 pilot test participants took two or three of the three versions in different orders in the course of an academic year. The test results indicated high reliability of the modified C-Test with the target group as well as high inter-version correlations. Equivalency among versions was investigated through analyses of the test results in relation to (1) item facility figures, (2) the pilot test participants, and (3) the time of the test administration during the academic year. The lower mean score for Version 2 was revealed to be partly due to lower item facility figures or a smaller sample, namely the fact that some of the high-scoring participants did not take Version 2.

Regardless of the order in which the three versions were presented to the participants in the course of the year, test scores enhanced invariably toward the end of the year. Although some degree of practice or memorization effect was suspected, no valid means was conceivable to separate it from the extent to which proficiency improvement was reflected in the mean scores. However, the score enhancement in all different orders of administration was considered sufficient evidence pointing toward inter-version equivalency.

Notes

1 The second of the alternatives has been realized subsequently, too. This alternative will eventually be reported in a future paper.

2 Hastings (1996 on website) claims that the three C-Tests developed at the University of Wisconsin-Milwaukee out of 12 passages with 10 blanks each are "constructed correctly and validated against dependable criteria."

3 What Hata calls "modified C-test" leaves first two sentences of a passage intact and deletes either the first or the last part of every other syntactic phrase. This "modification" of the C-test procedure is apparently necessitated both by the syntactic structure and the orthographic convention of the Japanese language.

4 In Klein-Braley and Raatz (1984), "it is suggested that the first text should be very easy and that the difficulty should increase throughout the test so that the final text is very difficult." (p.144) This is perhaps based on the premise that the test is made longer than any of the test takers can complete within the given test time.

5 This point adds to the evidence for construct validity of the C-Test, namely the validity of the testing procedure used in the C-Test, or more specifically, the modified C-Test. Incidentally, the previous research papers reviewed in the "Background" section of this paper compared C-Tests with various proficiency tests, but none with MEPT. The comparison between the modified C-Test and MEPT is to be reported in detail in Hiser, Ishihara and Okada (forthcoming).

6 Six out of eleven native speaking informants also gave "focus" instead of the expected "form."

7 Klein-Braley and Raatz (1984) also states that "Ideally the target group should score on average 50 per cent." (p.144)

8 The participant who scored highest, i. e. 98 points out of 100, missed only one item in Version 3, i. e. No. 46 "one." This item was correctly restored by 7.5% of the participants, while none of the native-speaking informants had difficulty with this item. The item facility figures for this particular blank were 0.223 for the top group and 0.036 for the low group; consequently the item discrimination 0.188 was relatively low due to the fact that this test item was difficult for the top group as well as for the low group. Besides, the student in question happened to take the C-Test four times: Version 1 (79 points) and Version 2 (81 points) in May, and Version 2 a second time in January (84 points), so that Version 3 in January (98 points) was the fourth time the same passages were presented to this participant.

9 Since the rescored totals varied for the three versions (74 points, 80 points, and 62 points for Versions 1 through 3 respectively), the means for the rescored totals were converted into percentage to make comparison easier.

References

Carroll, J. B. (1987). Review of "C-Tests in der praxis" by Klein-Braley & Raatz (1985). *Language Testing, 4*, 99-106.

Chapelle, C. A. & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7 (2)*, 121-146.

Cohen, A. D., Segal, M. & Bar-Siman-Tov, R. (1984). The C-Test in Hebrew. *Language Testing, 1*, 221-225.

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

Grotjahn, R. (www.slf.ruhr-uni-bochum.de/biblio/ctbib.html). The electronic C-Test bibliography.

Hastings, A. J. (www.su.edu/sas/tesol/indefens.htm). In defense of C-Testing.

Hata, K. (1990). 日本語クローズ・テストから日本語変形 C-Testへ：採点法の問題を中心に (From cloze to modified C-Test for the Japanese Language. 名古屋大学総合言語センター言語文化論集11 (2): 213-225. Also available at http://langnagoya-u.ac.jp/proj/genbunronshu/ronshu2.html.

Hiser, E., Ishihara, K. & Okada, T. (to appear). C-test validation for Japanese EFL students.

Ikeguchi, C. (1998). Do different C-tests discriminate proficiency levels of EL2 learners? *JALT Testing & Evaluation SIG Newsletter, 2 (2)*, 3-8.

Ishihara, K., Okada, T. & Matsui, S. (1999). English vocabulary recognition and production: A preliminary survey report. *Doshisha Studies in Language and Culture, 2-1*, 143-175.

Ishihara, K., Okada, T. & Matsui, S. (2000). Vocabulary levels analysis: Survey results with university students. *Doshisha Studies in Language and Culture, 3-1*, 17-46.

Jonz, J. (1991). Cloze item types and second language comprehension. *Language Testing, 8*, 1-22.

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

Kakkota, V. (1988/98?). Letter-deletion procedure: a flexible way of reducing test redundancy. *Language Testing, 5 (1)*, 115-119.

Kamimoto, T. (1992). An inquiry into what a C-Test measures. *Fukuoka Women's Junior College Studies, 44*, 67-79.

Kamimoto, T. (1993). Tailoring the test to fit the students: Improvement of the C-Test through classical item analysis. *Language Laboratory, 30* (November), 47-61.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing, 2 (1)*, 76-104.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 4 (1)*, 47-84.

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

Mochizuki, A. (1994). C-Tests: Four kinds of texts, their reliability and validity. *JALT Journal, 16 (1)*, 41:54.

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

Raatz, U. & Klein-Braley, C. (1996). Introduction to language testing and C-tests. In Coleman (Ed.). *University language testing and the C-Test*. (University of Portsmouth Occasional Papers in Linguistics). Also at http://uni-duisburg.de/FB3/ANGLING/FORSCHUNG/HOWTODO.HTM

Tsuchiya, T. (1998). What makes reading a difficult task?: The implications of a metrical analysis of college-level reading materials. *JACET Bulletin, 29*, 193-206.

Wiegle, S. C. (2000). Test review: The Michigan English language assessment battery (MELAB). *Language Testing, 17(4)*, 449-455.

Writing Research Group, JACET Kansai Chapter. (Ed.). (1995). Joint project report. In *Daigaku ni okeru eisakubun shidoo no arikata: Eisakubun jittai choosa no hookoku (Teaching writing at colleges and universities: A survey report*, 3-12.

Writing Research Group, JACET Kansai Chapter. (Ed.). (1997, 1998, 1999a). *Daigaku ni okeru eisakubun shidoo no kadai: Jissen kenkyuu no hookoku (Teaching writing in colleges and universities: Practical reports), 2, 3* and *4*.

Writing Research Group, JACET Kansai Chapter. (1999b). Two research projects. In JACET Kansai Chapter. 1999. (Ed.). *JACET Kansai-Shibu Kiyo, No. 5*, 51-55.

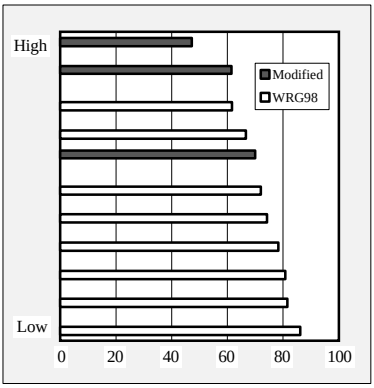
Appendix 1: Readability data for the three passages of the modified C-Test

(1) Flesch Reading Ease(*) and Flesch Kincaid Grade Level (**) for the three passages

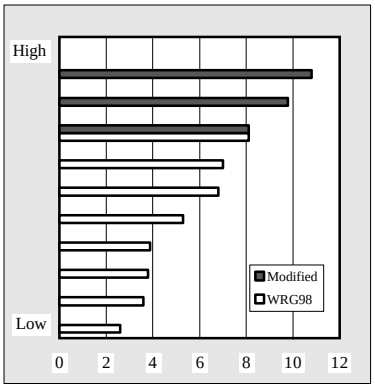
Passages	1. Public Alert	2. Advertisement	3. Space Shuttle	Average
Counts				
Words	112	120	143	125.0
Sentences	6	6	7	6.3
Averages				
Words/Sentence	18.7	20.0	20.7	19.7
Letters/Word	4.4	5.1	4.6	4.7
Readability				
FRE*	70.0	47.3	61.4	59.6
FKGL**	8.1	10.8	9.8	9.6

(2) Readability data of the three passages of the modified C-Test (solid bars) in comparison with those of the four passages of C-Tests 1 and 2 each (white bars) devised by the Writing Research Group of the JACET Kansai Chapter (1999)

A:Flesch Reading Ease



B:Flesch Kincaid Grade Level



Appendix 2: The test passages of the modified C-Test (The brackets indicate the deleted parts in Version 1 of the modified C-Test.)

Passage 1: Test passage "Public Alert"
(Based on the reading material "Police Description" in *Meanings into Words* by Dough, Jones & Mitchell. Cambridge University Press, 1984, p. 16)
Police are looking for a man in connection with this morning's bank robbery in Leicester. It is known that the sus[pect] is a man in his ea[rly] thirties, is lightly built, and i[s] about five feet eight inches ta[ll]. He has small eyes a[nd] a pale complexion with shoulder len[gth] brown hair. He is well dre[ssed], wears a gold ring on h[is] left hand, and speaks wi[th] a London accent. Police believe h[e] is still carrying the gun us[ed] in the robbery, and members o[f] the public are warned not t[o] approach him but instead to not[ify] the police immediately if he is sig[hted]. Extreme caution is urged in approaching the suspect.

Passage 2: Test passage "Advertisement"
(Based on the reading material "The Ultimate Advertising Medium" in *Academically Speaking* by Kayfetz & Spice, Hineley & Hineley, 1987, p. 109)
Radio remains a vital force in advertising, but television dominates the media world today. It is only natural that television has bec[ome] the dominant advertising medium as we[ll]. An important lesson that was fi[rst] learned about advertising on radio w[as] applicable to television also; in a mar[ket] flooded with numerous products, the fo[rm] of the ad was a[t] least as important as the con[tent]. When advertising on television began, i[t] was a challenge since adver[tisers] could now picture the product a[s]

well as describe it in wo[rds]. Cigarette commercials in the m[id]-1950s showed scene after scene o[f] spring fields. Clearly the mes[sage] was that smoking is like a springtime experience, embodying all the joys of youth, love, and picnics.

Passage 3: Test passage "Space Shuttle"

(Based on the reading material "The Shuttle and Beyond" in *Meanings into Words* by Dough, Jones & Mitchell, Cambridge University Press, 1984, p. 140)

The development of a space shuttle has dramatically reduced the cost of sending loads into space. The shu[ttle] is a reusable type o[f] space craft which takes o[ff] from the earth like a roc[ket], and lands like an airc[raft]. It can transport not on[ly] its own crew, but al[so] passengers, and has a hu[ge] cargo-hold which is cap[able] of carrying large satellites o[r] a space laboratory. It i[s] difficult to imagine the imm[ense] opportunities created by the shu[ttle]. One of the great advan[tages] of having a reusable sp[ace] vehicle is that it c[an] take one load after ano[ther] into orbit. Very large sp[ace] stations could not be laun[ched] in their complete form dire[ctly] from the earth, but they could be built piece by piece in space. The space shuttle is likely to be used as a general workhorse for the rest of this century.

Appendix 3: Parts of speech or sentential function of C-Test items

Version & Passage	Part of Speech	Function	Frequency	Cumulative Frequency	% of Version
1.1	Noun / pronoun	subject, possessive	3	03	6
	Verb, Verbals	main, descriptor, infinitive	5	05	10
	Adjective	descriptor, complement	3	03	6
	Preposition	phrase, infinitive	3	03	6
	Conjunction	connect nouns	1	01	2
1.2	Noun / pronoun	subject, object	7	10	20
	Verb, Verbals	main	2	07	14
	Adjective & Adverb	descriptor, comparison	3	06	12
	Preposition	phrase	2	05	10
	Conjunction	introduces comparative	1	02	4
1.3	Noun / pronoun	subject, object	6	16	32
	Verb, Verbals	main, modal	3	10	20
	Adjective & Adverb	descriptor, complement	7	13	26
	Preposition	phrase	1	06	12
	Conjunction	connecting clauses, nouns.	3	05	10
				Total	100
2.1	Noun / pronoun	complement, object, subject	6	06	12
	Verb, Verbals	main	3	03	6
	Adjective & Adverb	descriptor, complement	4	04	8
	Preposition	phrase, infinitive	2	02	4
2.2	Noun / pronoun	complement, object	2	08	16
	Verb, Verbals	main, descriptor	3	06	12
	Adjective & Adverb	descriptor	6	10	20
	Preposition	phrase	3	05	10
	Conjunction	comparative	1	01	2
2.3	Noun / pronoun	object, subject	6	14	28
	Verb, Verbals	main, modal, descriptor, gerund			
		Infinitive, auxiliary	6	12	24
	Adjective	descriptor	3	13	26
	Preposition	phrase	2	07	14
	Conjunction	connects clauses	3	04	8
				Total	100
3.1	Noun / pronoun	subject, object	4	04	8
	Verb, Verbals	main, auxiliary, complement	3	03	6
	Adjective & Adverb	descriptor	4	04	8
	Preposition	phrase	2	02	4
	Conjunction	connects clauses	2	02	4
3.2	Noun / pronoun	object, subject	3	07	14
	Verb, Verbals	main, gerund, descriptor	3	06	12
	Adjective & Adverb	complement, descriptor	3	07	14
	Preposition	phrase	4	06	12
	Conjunction	connects verb, clauses	2	04	8
3.3	Noun / pronoun	subject, object, possession	7	14	28
	Verb, Verbals	modal, main, gerund, descriptor	5	11	22
	Adjective	descriptors	3	10	20
	Preposition	phrase	4	10	20
	Conjunction	connects verbs	1	05	10
				Total	100

C テスト変形の試み

C テストは有効かつ総合的な外国語能力テストとして1981年にRaatzとKlein-Braleyによって提唱され、以来、今日まで多くの試行・応用と論議がなされてきた。1993年に大学英語教育学会関西支部のライティング指導研究会によって作られた C テストは、本学を含め関西地区のいくつもの大学で、毎年繰り返し調査や研究の一環として用いられてきたが、同志社大学の全受講者の英語力をよりよく反映するためには、新たなCテストの必要が感じられた。実情に即した新しい C テストの作成に際して、RaatzとKlein-Braleyの提唱したCテストの原則を、いくつかの点で変更する必要に迫られた。C テストに費やす時間を15分程度と規定し、文面の難易度（Readability）を従来のものより高くするために、やや長い文面を用い、話題の数を制限するなどの変更を加えたが、最も大きな相違は、テスト文面中（RaatzとKlein-Braleyの提唱した２単語おきでなく）５単語おきに単語の後半を埋める方式を取ったことである。その結果、同一文面を元にして三種の「変形 C テスト」が作られた。本来、C テストではテスト用文面の難易度がテストそのものの難易度を左右するとされてきたが、同一文面を用いた三種の C テストは文面の影響を排して、テスト語のみの難易を試す好条件を提供することになる。本稿は「変形 C テスト」三種間の同等性を中心に、テストとしての有効性と信頼性を検証し論考することを通じて、C テストのあり方に一石を投じるものである。