
Testing the Test: Comparing SEMAC and Exact Word Scoring on the Selective Deletion Cloze

GREGORY HADLEY

Nagaoka National College of Technology, Japan

JOHN NAAYKENS

Niigata University, Japan

A number of questions surround the issue of cloze tests and their various scoring methods, with proponents on both sides making contentious claims to their validity or lack thereof. Before language teachers can determine which side they agree with, they should investigate the matter themselves and make an informed decision based upon their findings. This paper studies two issues: 1) The statistical correlation between Exact Word and Semantically Acceptable (SEMAC) scoring methods on a selective deletion cloze test and 2) The reliability coefficients of the selective deletion cloze test using both scoring methods under a variety of classroom conditions. In a series of four experiments, Exact Word and SEMAC scoring were found to be correlated very highly, suggesting that both scoring methods are measuring the same language quality. Depending upon the pedagogic concerns of the teacher, either scoring method can be used. Care must be taken, however, as this study also finds the reliability coefficients of SEMAC scoring to be significantly lower than Exact Word scoring.

1.0 INTRODUCTION

There have been traditionally two distinct methods for evaluating cloze tests: Exact Word scoring and the Semantically Acceptable scoring method (also known as SEMAC). For several years there has been some debate about the validity of these scoring systems. For example, Ikeguchi (1995) and Owen *et al.* (1996), state that when exact word and SEMAC scoring systems are compared, there is almost always a very high correlation between the cloze test scores -- usually up to +.90 to +.95. Owen explains:

When the correlation between scores on two tests is this high, it suggests that the two tests are measuring the same thing. In the present case the very high correlations suggest that the two scoring systems are giving us the same information, are measuring the same quality (p. 41).

On the other hand, some researchers (cf. Alderson 1979; Klein-Braley 1981) claim that neither the Exact Word or SEMAC method are very reliable, since most cloze test designs themselves are flawed. Klein-Braley and Raatz (1984) assert that:

- Scorers do not agree with individual solutions on SEMAC scoring;
- Exact Word scoring is frustrating for learners and scorers alike;
- Correlations between two cloze tests often could not be demonstrated in their studies (p. 135).

Because there is so much contention about cloze test scoring methods, language teachers are faced with a number of decisions: They can choose to ignore the conflict, choose to side with whoever's argument seems to "make sense", or teachers can investigate the matter themselves and make an informed decision based upon their findings.

1.1 PURPOSE

Following a discussion of the background of cloze tests, this paper will discuss Exact Word and SEMAC scoring, and investigate whether or not there is a tendency for high correlations between the two systems. A recent study conducted with Japanese EFL learners will be reviewed in this paper. The implications this study has for the current testing debate, as well as for classroom teacher concerns, will be examined at the conclusion of this paper.

1.2 THE CLOZE TEST: AN OVERVIEW

Cloze testing was first introduced by W.L. Taylor (1953), who developed it as a reading test for native speakers. He defined the term "cloze" from a gestalt concept which teaches that an individual will be able to complete a task only after its pattern has been discerned:

A cloze unit may be defined as: any single occurrence of a successful attempt to reproduce accurately a part deleted from a 'message' (any language product), by deciding from the context that remains, what the missing part should be (p. 416).

Cloze tests consist of a text (usually two or three paragraphs) which has had words or parts of words deleted from it. Students or test subjects must then draw upon their knowledge of the language to write words which appropriately fill in the blanks.

There are at least five main types of cloze tests available to language teachers: The fixed-rate deletion, the selective deletion (also known as the rational cloze), the multiple-choice cloze, the cloze elide and the C-test (Ikeguchi 1995; Weir 1990; Klein-Braley and Raatz 1984).

In the fixed-rate deletion, after one or two sentences, every *n*th word is deleted. Usually every fifth or seventh word is deleted, but Brown (1983) suggests that longer texts with every eleventh or fifteenth word deleted can be used with subjects who have a lower level of language proficiency. In the selective deletion or rational cloze, the tester chooses which items he or she wishes to delete from the text. This allows teachers to fine tune the level of difficulty of the text, as well as define the test's pedagogic focus. Multiple choice cloze tests provide the subjects with several possible items to choose from for each blank in the cloze test. The cloze elide inserts words which do not belong in the text. It requires the subjects to identify the incorrect words and write in more appropriate items in their place. The C-test consists of deleting only part of every second word in a text, and involves subjects in completing each truncated word.

None of these cloze test options, as Hughes (1993) suggests, should be seen as the panacea for our testing needs. However, if care and pretesting are included in the process of making a test, then cloze tests can be very helpful general proficiency indicators of where our learners are in their process of acquiring the target language (cf. Brown 1991, 1988b; Chavez-Oller *et al.* 1985; Perkins and German 1985; Bachman 1982).

Before teachers can make such inferences on their learners' progress, they must decide how to score the tests. And before scoring the test, they must resolve for themselves which scoring system is right for their purposes.

1.3 EXACT WORD AND SEMAC SCORING

Except for the C-test and the multiple-choice cloze, most cloze tests use either the Exact Word or SEMAC scoring method. In the exact word method, the cloze test blanks are filled in by the subjects with the exact same word as was in the original text. Correct answers receive 1 point, while any other response receives no points. SEMAC scoring allows subjects to write answers which, though not the original words deleted from the text, are grammatically and lexically appropriate.

Many of the issues related to Exact Word and SEMAC scoring are dealt with in Owen *et al.* (1996, p. 40-42). They state that most teachers opt for SEMAC scoring, since they feel it is fairer to the subjects than with the Exact Word scoring method. However, SEMAC scoring can be much more difficult to apply, especially if foresight has not been used in making the cloze test. This is the major criticism of the method by Klein-Braley and Raatz (1984), Weir (1990) and Hughes (1993). Owen *et al.* (1996) maintain however that SEMAC and Exact Word scoring correlate strongly, giving teachers the choice to use either scoring system if they wish to, since both systems appear to be measuring the same language attributes.

Exact Word scoring tends to conceal the differences in ability between students, because answers that equally appropriate are not marked as correct, and the students are not affirmed for their collocational competence. For this reason, Owen *et al.* (1996) seem to fall on the side of SEMAC scoring, saying it is more internally-reliable than Exact Word scoring. They admit that this is a contentious claim, since cloze tests are rather organic in nature and probably shouldn't be measured by item facility (IF) statistics such as KR-20 (Griffiee 1995; Klein-Braley and Raatz 1984).

Do Exact Word and SEMAC scores correlate highly? Is the dispersion and reliability higher for SEMAC scoring? More importantly, what difference does all of this make for EFL teachers? These questions motivated us to conduct the following experiment in order to seek some possible answers to these questions.

2.0 METHOD

This experiment studied two issues: 1) The statistical correlation between Exact Word and SEMAC scoring methods on a selective deletion cloze test and 2) The reliability coefficients of the selective deletion cloze test using both scoring methods under a variety of classroom conditions. The study was conducted in the 1996 Fall semester at Niigata University, located on the northwestern coast of the main island of Honshu in Japan.

2.1 SUBJECTS

Two groups of subjects (see Table One) were selected for this study from Niigata University's First Year English 1B classes. All were native Japanese speakers from various

prefectures of the main island of Honshu. Group One consisted mostly of first year Science majors and two Elementary Education majors. Group Two was comprised of first year Engineering students.

Table One: Subjects		
	Group One	Group Two
Language	Japanese	Japanese
Age	18 (82%) 19 (18%)	18 (100%)
Sex	Male (55%) Female (45%)	Male (100%)
Department	Science (91%) Education (9%)	Engineering (100%)
Skill Level	False Beginners	False Beginners
Total Number Subjects	22	24

No special criteria was used in selecting or excluding the subjects. Neither group was tested on their English proficiency level before entering this course, except for the structuralist grammar-based entrance examination that all the subjects took a year before participating in this study. However, classroom experience with both groups led us to believe that most group members had limited speaking, listening and writing skills, typically representative of an EFL class of this level in a Japanese university setting (Wadden 1993).

2.2 MATERIALS

The learners were given the same selective deletion cloze on two different occasions (see Figures One and Two). The test adapted from a general interest reading text found in the first chapter of the course book (Richards et al. 1993, p. 7). While the subjects had read the text several months earlier, we were fairly certain that very few if not any had read the text again since that time. The cloze test consisted of a 133 word passage with 25 blanks, meaning that roughly 19% of the total text was deleted. This test was set up mainly as a criterion-referenced measure to help students assess for themselves whether or not they had adequately studied the key vocabulary, grammar and discourse elements which would be featured in the upcoming midterm exam. The first test was for the students' personal evaluation only, and the second was used as part of their grade for the course.

The selective deletion cloze is justifiable for this sort of evaluation. Bowen *et al.* (1985, p. 376) state that the selective deletion cloze is ideal for testing vocabulary and grammar. Bachman (1982, p 61-70) finds that the selective deletion cloze can be used to investigate a subject's knowledge of written discourse items such as context cohesion, syntax and strategic textual comprehension.

2.3 PROCEDURE

The cloze test was administered to both groups during their regular class period, and again during class two weeks later. On both occasions, the instructions were given to the students verbally and in written form, both in English and Japanese, to facilitate a clear understanding of the task. On each occasion, the cloze tests were collected after 20 minutes. According to Ikeguchi (1995), this is an acceptable amount of time to allow for Japanese college students to complete even short cloze tests. This allows for the "pre-testing strategies" often observed in Japanese students at this academic level:

Students look at the form of the test itself. Students then check to figure out if there is any "trick" to completing the test in a mechanical fashion without actually knowing the answer.

- If this cannot be established, the next step is to look for a puzzle-like consistency to the test.
- Once this has been investigated, then students begin tackling the actual test content and answer the questions.

This type of behavior has been re-enforced through years of testing and testing preparation in high school and *juku* -- private schools which teach how to pass university exams, and it is important to let them have time to go through this ritual. (Shimahara 1991; Fujita 1991; Tsukada 1991). One significant variable that was different, however, is that the first test was administered during a regular class session, while the other was given during their midterm test. While this is certainly not advisable when performing a test-retest experiment, we purposely attempted this in order to see how the test would react in under a variety of classroom conditions.

2.4 ANALYSIS

The tests for both groups were photocopied and graded by two scorers. The classroom teacher graded the tests using Exact Word scoring, while an expatriate TEFL lecturer who was unacquainted with the subjects graded the tests using the SEMAC method. We decided on this approach because SEMAC scoring often involves a subjective judgement on the subject's response. We did not want the SEMAC scores to be influenced one way or another by personal knowledge of the subjects. Before grading the tests, the blind marker was given a manuscript of the complete text, and instructed to allow any words in the cloze that were either synonymous, lexically and grammatically correct. Mistakes in historical accuracy, and minor spelling errors were to be ignored. If the scorer found himself in a situation where he had to think very hard as to if an answer was acceptable or not, he was free to mark it as incorrect.

After all the scores were figured, all of the data was analyzed using the VAR Grade for Windows 1.0 software package (Revie 1994). Often Kuder-Richardson formula 20 (KR-20) is used for single item tests where the reliability of a specific question can be tested (eg. multiple choice questions). While some have used KR-20 in testing the reliability of cloze tests, we take the position that the cloze test is much more of an organic test instrument whose parts cannot

be easily separated for valid analysis. Instead of using the KR-20, Test-retest was used in order to ascertain the reliability coefficients for the test using both scoring systems.

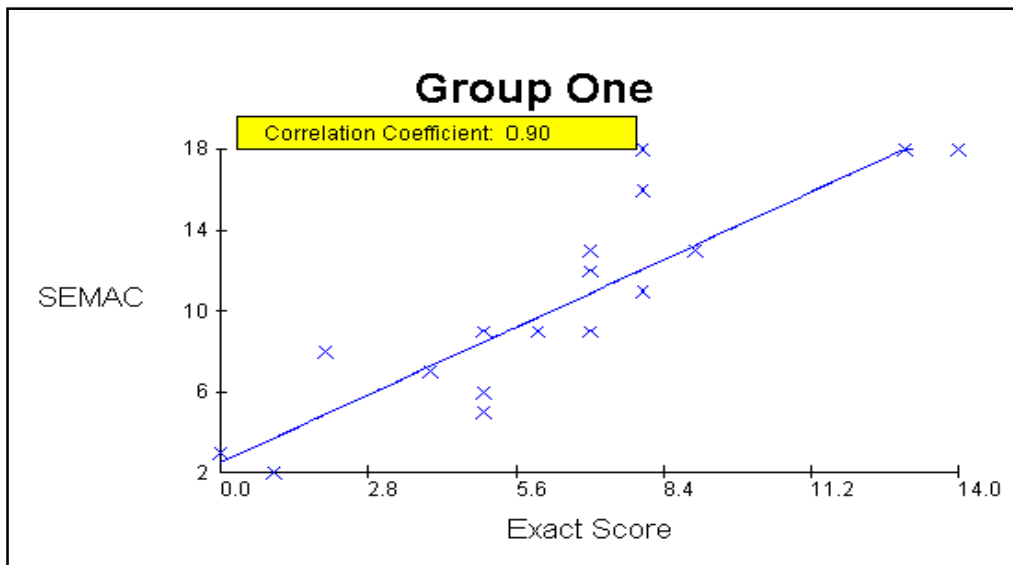


Figure 1

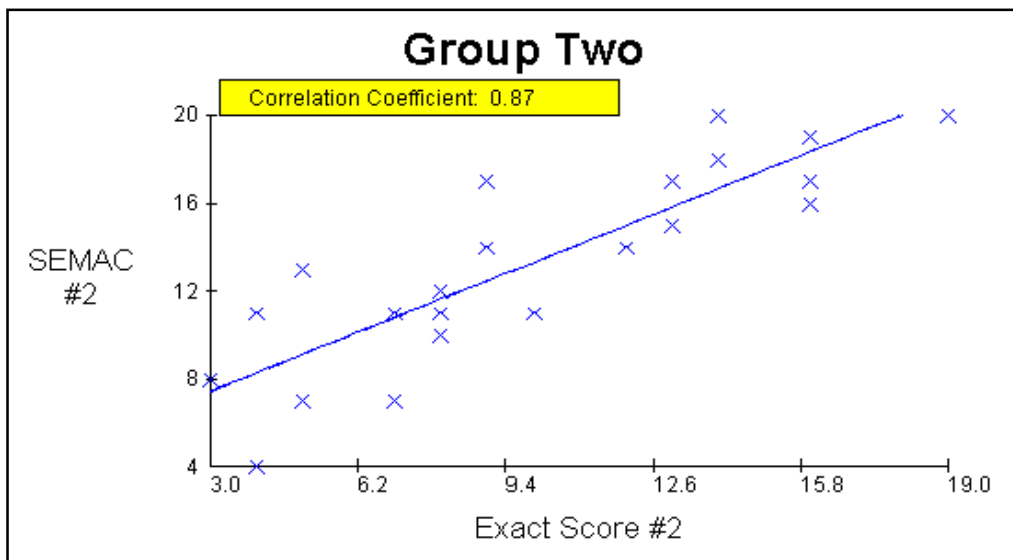


Figure 2

With Group One, the reliability co-efficient for the test using the Exact Word method was at +.60, while the reliability co-efficient for the SEMAC method was at +.56. Group Two's test-retest coefficients were considerably lower, with Exact Word at +.34, and SEMAC at +.31.

Correlating SEMAC and Exact Word scoring methods for Group One showed a correlation coefficient of +.90 on the first test and +.90 on the second test. Coincidentally, Group Two's correlation coefficient was +.87 for the first test, and +.87 on the second test (see Figures Three and Four).

3.0 DISCUSSION

It is true that the scoring methods were re-tested under very different conditions, because the first test was used as a diagnostic test, and the students were told that they would have a second opportunity to take the test again later for a grade. The higher means of the second tests (see Table Two) show that a students knew they hadn't studied enough, went back and searched for the text used in the cloze test, and most probably memorized it. Most did not take the time to study for the test even when they knew what was going to be on it. In our experience, many students feel confident that, once they have entered a university, studying is an option so long as they attend about 60% of their classes and get 60% on a final exam - an exam they will be allowed to take several times until they finally pass. Nevertheless, it would be interesting to administer the same cloze test after a period of two months to the same students and see if the scores correlate more strongly with the first test scores.

Table Two: Descriptive Statistics for Both Groups							
	Task Name	Students	High Score	Low Score	Mean	Median	Standard Deviation
Group One	Exact Score	22	14	0	6.6	6.5	3.7
Group One	Exact Score #2	22	21	6	13.0	12.0	4.7
Group One	SEMAC	22	18	2	10.4	9.0	4.9
Group One	SEMAC #2	22	23	6	15.9	16.0	4.7
Group Two	Exact Word	24	14	2	5.5	5.0	3.0
Group Two	Exact Word #2	24	19	3	10.1	9.0	4.5
Group Two	SEMAC	24	18	3	8.3	7.0	4.1
Group Two	SEMAC #2	24	20	4	13.4	13.5	4.4

Yet even when taking into consideration all of these normal classroom variables, we can see that the selective-deletion cloze is a robust and reliable measure. It is probable that if other tests were subjected to the same sort of abuse as in this study, they would not come out with reliability coefficients as high as between +.31 to +.60.

4.0 IMPLICATIONS FOR TEACHERS

If it is simply a question of wanting to know which scoring system is more reliable for cloze testing, this study suggests that either Exact Word or SEMAC will do. As a quick diagnostic tool, Exact Word is quicker and easier to score than SEMAC. If the cloze test is used as a C-RT, then SEMAC scoring gives students more points that can be applied to their grade.

However, the convenience to the teacher cannot be the only factor in determining which scoring system to use. Teachers may need to seriously consider how the advantages and disadvantages of either system might complement or clash with the values of their students' culture of learning. For example, in Japan we have observed that many students prefer the Exact Word system, because it gives the impression that there is one and only one correct answer. If given the complete text after taking the test, some might even try to memorize the words which were removed from the text, although often out of context and devoid of any real understanding on how to use the word. SEMAC scoring can be distrusted by students who resist the possibility for successful communicative variation in test answers. We are given the impression from some students that SEMAC scoring leaves them at the mercy of an arbitrary standard determined by the teacher.

The effect of the actual test scores on students in Japan raises other considerations. International observers have noted that Japanese society has tried unsuccessfully to create a system of Capitalist Socialism, which often saves the mediocre rather than rewards the talented (Roche 1999). This has seeped into the educational system as well, where there is a student who receives 60% and the student who receive 90% on the a test are treated the same. If the cloze test is used as an N-RT, then our experience has told us that our Japanese learners are more concerned as to how close they fit in the middle of the group, rather than how much better or worse their grade is from the median.

This of course is our experience with learners in Japan. Teachers in other countries may encounter different responses to tests based upon their students' culture. There will not be easy answers to these issues, only informed decisions based upon careful consideration of the students' needs as language learners. Keeping these concerns in mind may allow language teachers the confidence to conservatively try cloze testing in their own unique classroom environment.

5.0 CONCLUSION

This study stimulates a number of new questions, the answers to which would be valuable to classroom teachers. For example, this study relied upon one blind scorer to use the SEMAC scoring method. What would the inter-marker reliability be if the results of several scorers using the SEMAC method were compared on the same group of subjects? Or, how would the selective-deletion cloze compare with a C-test on the same text with a group of Japanese students? A number of studies (Hansen and Stansfield 1981; Stansfield and Hansen 1983; Hansen 1984; Chapelle 1988) suggest that field independence is an important factor in performing well on tests such as the cloze. Do Asian students, as a result of Confucian-based education systems that value conformity and homogeneity, tend to be more field dependent or independent? How would that change the way we look at cloze testing (or testing in general) for Asian ELT? If selective-deletion cloze tests do measure the level of a subject's knowledge

of vocabulary and grammar, would students with higher scores on the selective deletion cloze test also score higher on tradition grammar-based written examinations? If they did correlate highly, would that suggest that cloze tests could be used in place of or alongside of traditional paper tests, thereby providing a measure which is fair to students and easier for teachers to grade?

We hope these questions will motivate others to begin testing studies of their own. Such research not only helps us as teachers to improve our tests by making them more a productive part of the language learning process, but also helps us all to improve as language teachers.

THE AUTHORS

Gregory Hadley is an Assistant Professor of English at the Nagaoka National College of Technology. John E. Naaykens is a part-time Lecturer and doctoral student at Niigata University.

REFERENCES

- Alderson, J.C. (1979). The cloze procedure and proficiency in English as a second language. *TESOL Quarterly*, 13, 219-226.
- Bachman, L. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, 16, 61-70.
- Bowen, J.D., Madsen H, and Hilferty, A. (1985). *TESOL: Techniques and Procedures*. Rowley, MA: Newbury House Publishers.
- Brown, J.D. (1983). A closer look at the cloze: Validity and reliability. In J.W. Oller, Jr. (Ed.) *Issues in Language Testing Research*. p. 237-250). Rowley, MA: Newbury House.
- Brown, J.D. (1988b). What makes a cloze item difficult? *University of Hawaii Working Papers in ESL*, 7(2), 17-39.
- Brown, J.D. (1991). What test characteristics predict human performance on cloze test items? In *the Proceedings of the Third Conference on Language Research in Japan* (p. 1-26). Urasa, Japan: International University of Japan.
- Brown, J.D. and Yamashita S. (Eds.) (1995). *Language Testing in Japan*. Tokyo: The Japan Association for Language Teaching.
- Chapelle, C. (1988). Field independence: A source of language test variance? *Language Testing* 5(1), 62-82.
- Chavez-Oller, M.A., Chihara, T., Weaver, K.A., and Oller, J.W. Jr. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning*, 5, 62-82.
- Finkelstein, B. Imamura, A. and Tobin, J. (Eds.) (1991). *Transcending Stereotypes: Discovering Japanese Culture and Education*. Yarmouth, Maine: Intercultural Press.
- Fujita, H. (1991). Education policy dilemmas as historic constructions. in B. Finkelstein, *et al.* (1991). *Transcending Stereotypes: Discovering Japanese Culture and Education* (p. 126-134). Yarmouth, MA: Intercultural Press.
- Griffiee, D. (1995). Criterion-referenced test construction and evaluation. in J.D. Brown and S. Yamashita (Eds.) (1995). *Language Testing in Japan* (p. 20-28). Tokyo: The Japan Association for Language Teaching.
- Hansen, J. and Stansfield, J. (1981). The relationship between field dependent-independent cognitive styles and foreign language achievement. *Language Learning* 31, 349-67.

- Hansen, L. (1984). Field dependence-independence and language testing: Evidence from six Pacific island cultures. *TESOL Quarterly* 18, 311-24.
- Hughes A. (1993). *Testing for Language Teachers*. New York:Cambridge University Press.
- Ikeguchi, C. (1995). Cloze testing options for the classroom. in J.D. Brown and S. Yamashita (Eds.) *Language Testing in Japan* (p. 166-178). Tokyo: The Japan Association for Language Teaching.
- Klein-Braley, C. and Raatz, U. (1984). A survey of research on the C-test." *Language Testing*, 1, 134-146.
- Klein-Braley, C. (1981). Empirical investigations of cloze tests: An examination of the validity of cloze tests as tests of general language proficiency in English for German university students. Duisburg, Doctoral Dissertation.
- Oller, J.W. Jr. (Ed.) (1983). *Issues in Language Testing Research*. Rowley, MA: Newbury House.
- Owen, C., Reeves, J. and Widener, S. (1996). *Testing*. Birmingham, UK: University of Birmingham Press.
- Perkins, J. and German, P. (1985). The effect of structure gain on different structural category deletions in a cloze test. Paper presented at the Midwest TESOL. Milwaukee, WI.
- Revie, D. (1994). *VAR Grade for Windows: Grading Tools for Teachers*. Thousand Oaks, CA: VARed Software.
- Richards, J., Hull, J., and Proctor, S. (1993). *Interchange 2: English for International Communication*. New York: Cambridge University Press.
- Roche, D. (1999). Going nowhere fast. *Time Magazine*, August 9, p. 23.
- Shimahara, N. (1991). Examination rituals and group life. in B. Finkelstein, *et al.* (1991). *Transcending Stereotypes: Discovering Japanese Culture and Education* (p. 126-134). Yarmouth, MA: Intercultural Press.
- Stansfield, C. and Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly* 17, 29-38.
- Taylor, W.L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.
- Tsukada, M. (1991). Student perspectives on *juku*, *yobik_*, and the examination system. in B. Finkelstein, *et al.* (1991). *Transcending Stereotypes: Discovering Japanese Culture and Education* (p. 126-134). Yarmouth, MA: Intercultural Press.
- Wadden. P. (Ed.) (1992). *A Handbook for Teaching English at Japanese Colleges and Universities*. New York: Oxford University Press.
- Weir, C. (1990). *Communicative Language Testing*. Hemel Hempstead: Prentice Hall International Ltd.