International Journal of Assessment and Evaluation in Education
Vol 1/Dec 2011 (77-93) ISSN 2232-1926

The Use of Cloze Test to Determine The Proficiency Level of ESL Students

Jeanette Lim
Sunway University

Tunku Mohani Tunku Mohtar
Universiti Pendidikan Sultan Idris

Abstract

A well-constructed language test will yield useful information for admission, student placement, course evaluation and other concerns in language teaching, depending on the test objectives. This paper describes the development of a viable test instrument which can be used to determine the English language proficiency level of students enrolled in a Malaysian private university. Students' responses to the multiple-choice cloze tests are examined to ascertain the impact of gender, nationality and English language qualifications on the test scores. The findings suggest that the tailored cloze tests are quite reliable, showing a fairly strong relationship between the students' EPT scores and their English language qualifications; and for some of them, their achievement test scores. Although this study is a preliminary exploration into a subject that needs further probing, the test construction procedures may serve as a useful alternative to institutions which need a language proficiency test that is easy to administer and quick to grade.

Keywords *language testing, cloze procedures, multiple-choice, proficiency, Second Language (L2)*

INTRODUCTION

A well-constructed language test will yield useful information for admission, student placement, course evaluation, and other concerns in language teaching, depending on the test objectives. For high-stake tests like a post-admission English proficiency test, apart from test reliability and validity, ease of test administration and marking are important criteria. Being an integrative test which is objective and fairly easy to mark, the cloze procedure appears to be an attractive option. However, cloze tests are not without pitfalls. For example, with the nth-word deletion, some gaps may be difficult to fill and the deleted items may not test the language aspects the test writer is interested in (Brown, 2003). James Dean Brown who spent 25 years researching on cloze tests believes that a cloze test must be put through a tailoring process involving item analysis techniques for it to work well. In this study, the procedures he recommends are modified to construct an English Proficiency Test (EPT) which includes four multiple-choice (MC) cloze passages. The findings suggest that despite some limitations, the tailored cloze tests can be a reliable and useful test instrument for busy teachers and newcomers in language testing.

In this study, a second language (L2) means any language that is acquired after the speaker has mastered his or her mother tongue or first language. This means the speaker may be able to speak more than one second language. This broad definition suits the study context better, being a Malaysian private university with a sizable population of international students. For many of our students, English is just one of the languages they speak apart from their mother tongue.

Proficiency testing

Proficiency tests are designed to measure test-takers' ability "in certain aspects of the language without referring to any training or instruction they have had before in that particular language" (Hassan, 2002, p. 23). The purpose of such a test is to find out if the test-takers have sufficient command of the target language to perform a duty or commence a particular programme. It is common for tertiary institutions to construct their own proficiency tests because many of them do not have a single English language entry requirement. This is the situation we face in Malaysia as well. Most of our international students are admitted based on their high school results. Very few of them come with scores of established English proficiency tests like the International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL). Although English is the medium of instruction of our university, it is not our mother tongue. With most of our international students coming from non-English speaking developing countries, we need to set appropriate proficiency levels for entry requirement. Graham (1987) suggested that language entry requirement could be adjusted according to the types of ESL support service available to the institution. At Sunway University, English for specific academic purposes courses are built into the curriculum of different disciplines. There is also a pre-sessional intensive English language programme for students who do not meet the language entry requirement. So the EPT we develop should discriminate between students who can commence their degree or diploma programme, and those who need to join the pre-sessional English course.

The construct of L2 proficiency is related to "knowledge of language and ability to use the language in different modes (speaking, listening, reading, writing) in contextually appropriate ways" (Yamashita, 2002, p. 83). Second language proficiency models such as Canale's (1983) model of communicative competence and Bachman's (1990) framework of communicative language ability include knowledge of vocabulary and grammatical structure, which are related to reading comprehension (Yamashita, 2002, p. 83). Depending on how the test is constructed, cloze tests "contain various types of items requiring different levels of linguistic and cognitive processing" (Yamashita, 2003, p. 268). Cloze procedure "may be a test of grammar, of vocabulary, of discourse-level processing, or of general proficiency" based on the words deleted (Bailey, 1998, p. 71). Therefore, it seems a good choice as a proficiency test instrument.

Cloze tests

First introduced by Wilson Taylor in 1953 as a means to determine native speaker's reading ability, the classical cloze test "involves deleting every nth word from a prose passage" (Hinofotis, 1980, p.122) and test-takers are required to provide the missing words in the blanks. The nth-word deletion is also known as fixed-ratio deletion; the deletion rate could be every 5th, 7th, 9th or 11th word. Subsequent modified cloze procedures include rational deletion, where items measuring selected language traits are deleted, and C-test, where every second half of every second word is deleted (Chappelle & Abraham, 1990, p.122). Open cloze tests can be scored using the exact word or acceptable word method, while multiple-choice cloze test is scored discretely.

In the 1970's and 1980's, the potential of the cloze procedure as a proficiency test for ESL learners captured many researchers' attention, especially when earlier studies on cloze tests show high correlations between cloze test results and total scores of established ESL proficiency measures such as the TOEFL (Brown, 2002; Oller, 1972; Hinofotis, 1980). Much of what was done on cloze tests during this period focused on developing, scoring and interpreting cloze procedures to increase their reliability and validity (Brown, 2003, p. 77). For example, Hinofotis (1980) showed that test length, text difficulty and student ability will affect cloze reliability while Alderson (1979) was interested in the effect of variation in text difficulty, scoring procedure and deletion frequency "on the relationship of the cloze test to measures of proficiency in English as a Foreign Language (EFL)" (p. 219). His study indicates that "different texts may well measure different aspects of EFL proficiency, or the same aspect more efficiently or less efficiently" when the same deletion rate used for different texts leads to different correlations with the criterion (Alderson, 1979, pp. 222-223). Furthermore, a change in the deletion rate can change the cloze test validity drastically (Alderson, 1979, p. 224). The apparent sensitivity of cloze to word deletion prompted Alderson to suggest that it basically tests linguistic skills of a relatively low order. He also argued that "cloze is essentially sentence-bound" because as a test, it is "largely confined to the immediate environment of a blank" (Alderson, 1979, p. 225).

However, Brown (2003) demonstrated in his 1989 study that, with the low level students who can only deal with sentence level grammar, cloze tests are likely to be testing mostly at the sentential level "because those will be the only items that are discriminating even marginally" (p. 84). Yet, with advanced students who can grapple with both sentence level grammar and inter-sentential level cohesion and coherence, cloze tests are likely to test both the sentential and inter-sentential levels "because those will be the items that are discriminate and therefore contribute to test variance" (Brown, 2003, p.84).

In another study, Brown (1984) contended that the most important variable in determining the reliability and validity of a cloze procedure could be "how well a given passage `fits' a given sample" (2003, p. 20). Variables such as scoring methods, deletion rate, blank lengths, text difficulty, test-takers and number of items seem to influence the `fit' of a particular cloze in varying degrees (Brown, 2003, p. 20). The extent to which a particular combination of these variables shapes the reliability and validity coefficients of a cloze test depends on how well these variables contribute towards the mean and standard deviation of the said test (pp. 20-21). Brown's study focus was "the interpretation of cloze tests for norm-referenced purposes" (1988, p. 21). He wanted to know the effect on the mean, standard deviation, validity coefficients of a cloze test when it was redesigned using item facility and discrimination indices as criteria for item selection (Brown, 1988, p. 21). Brown varied the deletion (every 7th word) starting point of an intermediate level text and gathered 250 different cloze items for the purpose of item analysis. 50 `best' items were then chosen based on the following criteria:

a. items with or closest to 50% in item facility, and
b. items with the highest discrimination indices. (1988, p. 23)

His results show that the tailored cloze is more reliable than the original cloze. This is quite remarkable because the test population consisted of students with a fairly narrow range of language proficiencies. The higher means and standard deviation of the tailored cloze are tested to be significantly different from those of the original cloze. This indicates that the item analysis techniques are effective in improving the reliability and validity of cloze tests although the "process is not as clear-cut and direct as that of a discrete point test" (Brown, 1988, p. 29).

Chappelle & Abraham (1990) wanted to find out "how variations in the cloze procedures affect measurement" in their study (p. 121). They constructed cloze tests based on the same text using four different procedures: fixed-ratio (every 11th word except for the first two sentences), rational, multiple-choice and C-test. For their multiple-choice version, the deletions were the same as the rational cloze except that test-takers were given options. As expected, the test-takers found the fixed-deletion cloze most difficult, followed by rational cloze, C-test, while the rational multiple-choice cloze was the easiest. The difference in their mean scores was significant but the reliability coefficient of each cloze procedure did not differ greatly. Since the only difference between the rational and multiple-choice cloze tests is the test format, this suggests that test format is a "determinant of difficulty level" (Chappelle & Abraham, 1990, p. 139).

Bachman's 1985 study addressed an interesting problem related to the validity of cloze tests. In order to find out what cloze test measured, he categorised the various kinds of language processing required to fill in any given blank in a cloze passage:

Type 1: information located within the clause;
Type 2: information spread across clauses but within a single sentence;
Type 3: information distributed across sentences but within the text; and
Type 4: extra-textual information." (Bachman, 1985, pp. 62-63)

He discovered that Type 1 and Type 4 gaps predominated in the 11th word fixed-ratio deletion cloze. This finding casts doubt on whether cloze tests actually measure discourse level processing (as represented by Type 2 and Type 3, above). However, Bachman (1985) suggested that more research using cloze passages based on different texts needed to be done in this area.

Although the findings regarding the reliability and validity of cloze tests remain inconclusive, the cloze procedure "often turns out to be a reasonably good test of overall English language proficiency" (Brown, 2002, p. 80). It is important to note that the cloze procedure is only "a technique for producing tests"; and "each test produced by the technique needs to be validated in its own right and modified accordingly" (Alderson, 1979, p. 226).

RESEARCH QUESTIONS

The research questions for this study are as follows:

1. Is there any difference in students' response to the multiple-choice (MC) cloze tests based on gender, nationality, and English language qualifications?
2. How reliable are the four MC cloze tests in determining the English proficiency level of the target population?
3. To what extent do item analysis techniques overcome the limitations of the MC cloze tests?

In Question 1, the objective is to find out if the students' gender, nationality and English language qualifications have any impact on their test scores. This concerns the issue of fairness – it is important to find out if the differences in the students' performance are "related primarily to the skills under assessment rather than to irrelevant factors" (*Code of Fair Testing Practices in Education*, 1988, in Kunnan, 2000, p.2). One of the aims of the study is to see if there is any correlation between the EPT scores and test-takers' English language qualifications. This should help the researcher gain a better understanding of the validity of the EPT scores, and decide if there is sufficient ground to make a recommendation to the admission office about which students need to undergo the pre-sessional English programme. Hence, the focus of Questions 2 and 3 is to see how the need for speed and reliability can be met without compromising test validity.

PARTICIPANTS

The participants of this research project were all students enrolled in the pre-university, diploma and degree programmes of our institution in 2006. Students from two pre-university programmes were involved in the test construction process. The subjects for the pilot test (58 of them) and the actual test (133) were degree and diploma students. Table 1 provides the composition of the latter in terms of gender and programme.

Table 1 Composition of actual test participants

Programme	Male	Female	Total
Diploma	59	23	82
Degree	37	14	51
Total	96	37	133

TEST FORMAT

For practical reasons, the multiple-choice (MC) format was adopted for our EPT. Firstly, the need for the EPT results to be produced speedily is of paramount importance. New students need to know if they could join the academic programme of their choice before the registration period closes. Secondly, the MC format should be familiar to most students since it has been around for a long time. Moreover, it is possible to construct MC items which require "a constructive, integrative process of comprehension … because the distractors used provide accurate information or inferences from the passage, but are incorrect answers to the questions (except for one appropriate answer to the question)" (Wixson & Peters, 1987, cited in Loh, 2006, p. 48). More importantly for this study, MC tests "enable easy and objective item analysis … and provide diagnostic information through analyses of responses to the alternatives in the items" (Loh, 2006, p. 47).

Loh (2006) recommends that test developers be guided by the following criteria when constructing MC items:

a. MC questions should be passage dependent so that they measure test-takers' comprehension of the passage rather than their background knowledge of the topic.
b. MC questions and responses should use test information from different levels of the passage, details as well as the main ideas so that the textual relationships are assessed, not just isolated pieces of information. (pp. 51-52)

When constructing MC cloze tests, it is possible to observe the criteria mentioned above because a cloze is "a method applied to a passage, and is therefore contextualized" (Steinman, 2002). In a fixed-ratio deletion cloze, "all classes and types of words have an equal chance of being deleted … making it a truly integrative reading activity" (Steinman, 2002). Hence, MC cloze tests can be a good choice when testing language proficiency.

The cloze procedure of every 5th word deletion is adopted in this study. A total of 25 words are deleted for each passage, except for the first paragraph. If the deleted word is particularly difficult, as in the case of an uncommon noun or a low frequency proper noun, the ratio is adjusted by changing the sentence structure or certain words or phrases without altering the meaning of the original sentence (Oller, 1983, cited in Steinman, 2002). Nevertheless, research has shown that the fixed ratio deletion cloze is more difficult for test-takers as compared to rational deletion (Chappelle & Abraham, 1990).

The 5th word deletion is used for the EPT because research has shown any deletion that is less than the 5th word will make the cloze `un-restorable' (Alderson, 1979). Also, since the passages chosen are fairly short, between 200 to 300 words long, the 5th word deletion is necessary in order to create 100 blanks or 100 MC cloze items. Earlier studies have shown that more MC items are needed to increase the internal reliability of the tests (Hinofotis, 1980; Alderson, 2000). This means more linguistic items are tested, both in frequency and types, and should thus reduce guessing on the test-takers' part while making the proficiency test more comprehensive.

Cloze passages

Bailey (1998) identified several variables which govern the difficulty level of a cloze passage:

> 1. the length of the text as a whole;
> 2. the amount of time the learners are allowed to complete the tasks;
> 3. the learners' familiarity with the vocabulary and the syntactic structures in the passage;
> 4. the length and complexity of the sentences in the text;
> 5. the learners' familiarity with the topic and with the discourse genre of the text (note that this item embodies the concepts of *content schemata* and *formal schemata*); and
> 6. the frequency with which the blanks are spaced (every fifth word versus every ninth word, for instance). (p. 62; *italics author's*)

The texts selected for the EPT are very much guided by these considerations. Four reading passages of various difficulty levels are taken from texts the target population is likely to encounter. The first two passages are intermediate level texts, adapted from a magazine for ESL learners. Passage 3 is a TOEFL text while Passage 4 is an IELTS text.

Test construction

The procedures for the construction of MC cloze test items and distractors are as follows:

a. Passages of the right level of difficulty are taken from various sources and turned into cloze tests.
b. For each passage, from the second paragraph onwards, every fifth word is deleted until a total of 25 blanks are created. The first paragraph is left intact to provide some degree of "conceptual build-up" (As qtd. in Oller, 1994, p. 384). Note that length of the first paragraph should not exceed 50% of the total length of each text. Otherwise test-takers may be sidetracked from or have less time for the tasks at hand.
c. The passages are "doctored" if the same word or linguistic item is deleted too many times, without changing the original substantially.
d. The passages were first converted into open cloze tests and tried out on diploma students of earlier intakes, as well as the students of the Intensive English Programme and two pre-university programmes .
e. The answers provided by the tryout groups were used to generate a list of distractors for each cloze item. For each item, note that the distractors should belong to the same word class as the deleted word.
f. The multiple-choice cloze tests were then given to the pilot groups.
g. Answer scripts were scanned through an Optical Marks Reader (OMR).

h. OMR analysis returned item analysis, test statistics, respondent statistics, and frequency distribution. Problematic items were identified and modified accordingly.
i. The revised test items were evaluated by the Language Centre staff of the university before being included in the actual test.

The tests were piloted on students with a wide range of language abilities so that there could be a variety of answers to be used as distractors for the cloze items. Such distractors could be considered more `authentic' because they reflected, to a certain extent, how the cloze items were perceived by the students through the kind of errors they made. Pre-university and IEP students were included also to help ascertain the difficulty level of the selected texts. The passages would be too difficult if the former could not attempt them since the proficiency level of our pre-university students is normally higher. Likewise, the passages would be too easy if the IEP students with a proficiency level equivalent to IELTS band 3 could answer easily.

Pilot and actual tests

The tryout subjects came from the programmes earmarked for the actual test. The test sessions occurred on different days but all were within two weeks of each other. The pilot groups were given one and a half hours to attempt the test but many of them completed it within an hour. Therefore, the testing time was shortened to one hour and 15 minutes during the actual test.

All the tests took place during regular class time under subject lecturers' supervision.

DATA ANALYSIS

The test scores generated were put through the following statistical analyses:

a. Item analysis for the objectively-scored cloze tests
b. Kuder-Richardson Formula 20 and Cronbach Alpha to ascertain the reliability of the EPT.
c. Inter-correlations between the EPT and the English language qualifications of the target groups to determine the concurrent validity of the EPT;
d. One-way Analysis of Variance (ANOVA), to compare the means of different groups on a particular variable.

Classical Item Analysis which requires the calculation of two measures – facility value (FV, to ascertain item difficulty) and discrimination index (DI, to determine how well the item discriminates between students of different proficiency levels) – was used. The FV of a good item is between +.20 (20%) and .80 (80%) while its DI should be equal to or greater than 0.20 (Garrett, 1966:368 as quoted in Castillo, pp.75-76). Although Bachman (1990) and Brown (2003) consider .20 - .70 to be an acceptable range for FV, the researcher settled for Garrett's recommendation as this is only a preliminary study.

Post-test Interviews

Six students were interviewed informally to find out what made the cloze tests easy or difficult for them. Three of them scored poorly in the cloze tests while the other three were the top scorers. The other questions focused on test-takers' familiarity with the test format, difficulty and interest levels of the texts as well as the students' test-taking strategies.

FINDINGS

The statistical analyses returned by the Optical Marks Reader (OMR) are presented in Table 2:

Table 2 EPT Descriptive & Inferential analyses

Statistic	Value
Number of Tests Graded	133
Number of Graded Items	100
Total Points Possible	100.00
Maximum Score	95.00
Minimum Score	27.00
Median Score	58.00
Range of Scores	68.00
Percentile (25)	45.00
Percentile (75)	67.50
Inter QuartileRange	22.50
Mean Score	58.32
Variance	218.43
Standard Deviation	14.78
Confidence Interval (1%)	55.02
Confidence Interval (5%)	55.81
Confidence Interval (95%)	60.84
Confidence Interval (99%)	61.62
Kuder-Richardson Formula 20	0.92
Coefficient (Cronbach) Alpha	0.92

The mean for the target population is 58.32 while the standard deviation is 14.78. Since the median score (58) and the mean are almost the same, this strongly suggests that the MC cloze tests are norm-referenced. It can be said that the EPT is a fair test since the mean score is more than half the total; and the standard deviation of 14.78 is wide enough for a norm-referenced proficiency test. This means the test is able to discriminate between the good and weak test-takers. A reliability score of .92 for Kuder-Richardson Formula 20 is a solid one, showing great consistency of the test items. The variance of 218.43 suggests that there is quite a bit of overlap, which is to be expected for a general proficiency test. The inter-quartile range is 22.5 and yet the score at the 25th percentile is 45, again indicating that this test is not too difficult for the majority of the target population.

Research question 1

Is there any difference in the students' response to the MC cloze tests based on gender, nationality, and English language qualifications?

Table 3 displays the cross tabulation between gender and their EPT scores:

Table 3 Cross tabulations between gender & EPT scores

Gender	EPT Score		Easy	Medium	Difficult
Male	Mean	55.98	28.83	22.65	4.53
	N	96	96	96	96
	Standard Deviation	14.24	5.99	7.79	2.47
Female	Mean	64.41	32.27	26.41	5.73
	N	37	37	37	37
	Standard Deviation	14.59	4.68	8.32	3.10
Total	Mean	58.32	29.79	23.69	4.86
	N	133	133	133	133
	Standard Deviation	14.78	5.85	8.09	2.70

Notice that there are more male than female students, mainly because the bulk of the test-takers (39 degree and 41 diploma levels) were students from the School of Computer Technology, which is traditionally male-dominated. The mean score of the female students, 64.4, is almost 10 points more than that of the male students (55.98). `Easy' refers to cloze items with a FV of .7 and above, `medium' refers to those items ranged between .36 and .69 while `difficult' refers to items ranged .35 and below. The mean scores of the female students for the medium and difficult sections are higher than those of the male students and the total target population. The one-tailed ANOVA test shows the score difference between the male and female students is significant. This suggests that the female students have better overall language ability. The reason could be that 10 out of the 37 female students were enrolled in the Bachelor of Psychology course which stipulates a higher English language entry requirement than the other degree programmes of our university. Their individual EPT scores are consistently high. More proficient test-takers, whether male or female, are expected to perform better in a language proficiency test. It is no exception in our context.

Most of the international students tested come from Indonesia (22). The other 29 students come from East Asia, South Asia, Africa, Middle-East and Europe. Their sample size is too small for any meaningful correlational studies, so they are divided into two categories: Indonesians and others. In a sense, this has limited the scope of the study since the researcher had hoped to gain some information regarding the international students' English language qualifications from the EPT. Table 4 shows a comparison between the test performance between Malaysian and international students**.**

Table 4 Cross-tabulations between nationality and EPT

Country	EPT Score		Easy	Medium	Difficult
Malaysia	Mean	60.37	31.56	24.56	5.24
	N	82	82	82	82
	Standard Deviation	15.90	5.78	8.69	3.07
Indonesia	Mean	54.82	28.64	22.36	3.82
	N	22	22	22	22
	Standard Deviation	11.76	5.10	7.01	1.94
Others	Mean	55.21	28.48	22.24	4.59
	N	29	29	29	29
	Standard Deviation	12.77	6.38	6.90	1.72
Total	Mean	58.32	29.79	23.69	4.87
	N	133	133	133	133
	Standard Deviation	14.78	5.85	8.09	2.71

The statistics indicate that the mean difference between Malaysian and international students is not significant although all the top scorers are Malaysians. A larger sample size for international students is needed for anything conclusive to be drawn. From the interviews conducted, two of the lowest scorers are international students and they claimed not to have done cloze tests before. This unfamiliarity may have contributed to the lower EPT scores for the international students; however, the impact of test format on their performance is yet to be determined.

The students' diverse English language qualifications are categorised into three main groups for the purpose of statistical analysis. Level 1 consists of English language qualifications deem acceptable for tertiary studies; Level 2 are qualifications considered sufficient for diploma programmes. Qualifications categorised under Level 3 are considered poor while unfamiliar language qualifications are grouped under Level 4.

As expected, students with better English language qualifications performed well in the EPT. (See Table 5) The mean score of holders of acceptable English language qualifications is much higher than the other two groups. The majority of these students are degree students, thus the EPT cut-off point could be 65 for those who wish to commence a degree programme at our university. Those with Level 2 English qualification have a mean of 55, the cut-off point could be between 50 and 55 for those who wish to commence diploma programmes. The mean score of those who come with poor English language qualifications is 18 points lower than the mean score of the total test population. Students who score lower than 50 should be advised to join the pre-sessional IEP. This shows that the EPT does discriminate between high and low proficiency students. The differences in these mean scores are tested to be significant.

Table 5 Cross-tabulations between English language qualifications EPT scores

English Language Qualifications		EPT Scores
Level 1	Mean	67.18
	N	51
	Std. Deviation	13.94
Level 2	Mean	55.35
	N	64
	Std. Deviation	12.04
Level 3	Mean	40.17
	N	12
	Std. Deviation	8.47
Total	Mean	58.66
	N	127
	Std. Deviation	14.94

Research question 2

How reliable are the four MC cloze tests in determining the English proficiency level of Sun-U students?

Reliability in testing means "dependability, in the sense that a reliable test can be depended on to produce very similar results in repeated uses" (Jones, 2001, p. 1 as qtd. in Weir, 2005, p. 22). Increasingly, reliability is seen as a kind of validity evidence following Alderson's example (1991).

The reliability of the EPT was calculated through the OMR statistical package. Two types of reliability coefficients are provided: Kuder-Richardson (KR) 20 and Cronbach's Alpha. Both are estimated to be .92, which meets the requirement of a high-stake test. However, the assumptions of KR 20 are that the test items are of equal difficulty level and independent of each other. For MC cloze tests, "being able to fill some gaps may depend on the ability to fill others, in which case test-takers' scores for filling these gaps are not independent of each other." Moreover, "the two halves obtained may not be parallel statistically" (Bachman, 2004, p. 161). Hence, he recommends that in calculating the reliability estimate of a test with interdependent items (such as cloze tests), "it may be preferable to split the test so that pairs of interdependence items are included in the same half" (p. 161). Hence, Guttman's estimate, which assumes item independence but not of equal difficulty, is calculated for the EPT as well. The cloze items are divided in two ways: the first half versus the second half, and odd number items versus even number items, as shown in Table 6:

Table 6 Guttman's split-half reliability estimates

Item	Guttman Split-half
1 – 50	0.8915
51 – 100	
Odd no.	0.92378
Even no.	

The first-half versus second-half calculation yields a slightly lower reliability coefficient than the odd-number versus even-number calculation, possibly due to the fact that the first two cloze passages are easier than the last two passages, so the test score range may not be as well spread out as the latter's. Nevertheless both can be considered to have met the requirement for a high-stake test.

Research question 3

To what extent do item analysis statistics and techniques overcome the limitations of multiple-choice cloze tests?

The method used in constructing the four MC cloze tests proves to be effective in that the test statistics for both the pilot and actual tests are almost the same. The mean for the pilot test is 58.29 while it is 58.32 for the actual test. As for standard deviation, it is 14.25 for the former and 14.78 for the latter. The overall FV of the pilot test is .58; it is the same for the actual test. The mean for the overall DI for the pilot test is .32, and for the latter, .33.

The high reliability estimates of 0.91 for both KR 20 and Cronbach's Alpha persuaded the researcher to make only minimal changes to the pilot test. Only those distractors that were not selected by anyone were changed and that involved only 8 items. For some items, their non-functioning distractors remained intact because distractors have to belong to the same word class as the answer, and it is difficult to find similar distractors. For example, the answer for item 44 is \`clicks', and the distractors are \`clicked', \`clicking' and \`click'. Since the linguistic item tested is verb form, it is difficult to replace \`clicked' as a distractor, even though no one chose it during the pilot test. It was decided then to keep the cloze items as they were, unless there were clear alternatives. It was also hoped that the non-functioning distractors in the pilot test would function when the test sample is larger. This proved to be the case for some items. The other items that did not function well for the target population should be revised later.

Such is the limitation of the MC cloze tests because there are guidelines for item construction. The distractors should be plausible, belonging to the same word class as the answers and matching the level of text difficulty, just to name a few. Moreover, the fixed-ratio deletion method is used to cloze the passages. There cannot be too much `doctoring' done; otherwise the original meaning of the texts would have altered. This may either increase or lower the text difficulty, or distort the original writing style. Brown (2003) actually recommends using the rational deletion method to cloze passages, which gives test developers greater control over what is being tested.

On the whole, the MC cloze passages have worked well for both pilot and target groups. It is effective to use the pre-university and diploma students' answers for the open cloze tests as distractors as the errors made are common for most students. Most of the distractors succeeded in distracting the other students, especially the weak ones. Item analysis statistics indicate that 22 items have a FV higher than .8 and only 3 have a FV below .2. In other words, 75% of the items are acceptable in term of FV. As for DI, 16 items are below .2, the acceptable range. This means 86% of the items discriminate well.

Since the MC cloze tests belong to the category of integrative test, it was hoped that they could replace the writing subtest of an earlier EPT. In Table 7, essay refers to the essay sub-test of the earlier EPT. Those who scored Band 5.5 and above for the essay sub-test were categorised as Level 1; and admitted into the degree programmes. Band 4 and 5 were categorised as Level 2; and those who were in this level would be admitted into the diploma programmes. Band 3.5 and below were categorized as Level 3; students placed in this level were expected to attend the pre-sessional IEP. Only the essay sub-test scores were correlated with the cloze test scores.

Table 7 Descriptive statistics for essay sub-test and MC cloze tests

Essay sub-test results		**EPT Score**
Level 1	Mean	78.56
	N	16
	Std. Deviation	12.65
Level 2	Mean	58.0
	N	54
	Std. Deviation	12.65
Level 3	Mean	49.38
	N	39
	Std. Deviation	12.25
Total	Mean	58.66
	N	109
	Std. Deviation	15.34

The statistical results show that students who did well for the essay sub-test also did well for the MC cloze tests, and vice versa. The ANOVA test confirms that the difference in these mean scores is significant. Nevertheless, this is not sufficient proof that the MC cloze tests can take the place of the essay sub-test. Firstly, not all the students sat for both the essay sub-test and the MC cloze tests. Secondly, a more detailed analysis indicating what both tests are testing is yet to be done. Furthermore, both inter-rater and intra-rater reliability need to be established for those who rated the essay sub-test, although they are all experienced English Language teachers.

Post-test interview

Post-test interviews were conducted for three top scorers and three lowest scorers to find out how they felt about the EPT. These interviews provide useful information on how they interacted with the test.

According to Brown (2003), the weaker students read the cloze passages word by word, sentence by sentence mainly because they could only handle sentence level grammar. This is confirmed by the interviews of the two weaker students. As expected, the good students were familiar with the test format. They made use of test-taking strategies when they took the test. For example, all three top scorers claimed to have previewed all the passages before attempting to answer the cloze items. When they were choosing the answers, they read through all the sentences with care and considered each of the distractors. One of them actually preferred to fill in the blanks before finding out what answers were available. This supports Brown's comment that higher proficiency students are able to handle both the sentential and intersentential grammar. Moreover, this seems to reinforce the view that cloze test is a good way to test reading. While the better students had ample time to complete the MC cloze tests, the weaker students struggled to comprehend the passages. The two international students did not have sufficient time to complete all the items.

Strengths of the MC cloze tests

There are several advantages in adopting the MC cloze tests as the EPT for our institution. Firstly, marking can be done through the OMR, which will return statistical test information for further analysis quickly and accurately. In its present state, the MC cloze tests already generated strong reliability coefficients. Once the non-functioning items have been removed or modified, the test statistics would be better. Secondly, the test method can be adapted to yield the sort of information test developers are looking for. Depending on stakeholders' requirements and the pedagogical context, the passages used can vary, so can the word deletion frequency and the marking method. The cut-off point for any particular measurement can be set fairly easily, as long as the criteria are spelled out. Thirdly, the test method combines the best of the multiple-choice format, yet can still serve as an integrative test. The validity of the MC cloze tests can be improved by adding an essay sub-test if they are to be used for proficiency testing.

CONCLUSION

The results of this study have disclosed some limitations of classical testing theory. The reliability of the test score is dependent on sample size (whether it is large enough for any meaningful conclusion to be drawn), composition (whether it reflects the target population), other than text type and test item construction. Test developers are aware that "the wider the ability range" of the test-takers, "the higher the reliability coefficient" (Weir, 2003, p. 31). It is easier to rank candidates of a wide range of abilities. This means that the test items may not be that well-constructed, but as long as they are given to a large enough number of subjects with diverse language abilities, the reliability coefficient is likely to be respectable.

There are a few things that should be done to improve the EPT further. One parameter Rasch model could be used to analyse the test items. The fundamental theorem guiding this model is that " an individual's expected performance on a particular test question, or item, is a function of both the level of difficulty of the item and the individual's level of ability" (Bachman, 1990, p. 203). This will rule out the weakness of sample dependency in classical testing theory, and the information produces through the test scores can be generalized to other settings as well, as long as the person's ability remains the same. The challenge is how to define the difficulty level of particular items.

REFERENCES

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

Alderson, J.C. (1980). Native and nonnative speaker performance on cloze tests. *LanguageLearning*, *30*, 59–76.

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

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

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bailey, K. (1998). *Learning about language assessment*. U.S.A.: Heinle & Heinle.

Brown, J.D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing*, *5*, 19–31.

Brown, J.D. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies, 21,* 79–125.

Brown, J.D. (2003). Twenty-five years of cloze testing research: So what? In Gloria Poedjosoedarmo (Ed.), *Teaching and assessing language proficiency: Anthology series 45* (pp. 46–111). Singapore: SEAMEO Regional Language Centre.

Castillo, E.S. (1990). Validation of the RELC test of proficiency in English for Academic Purposes. *RELC Journal, 21*, 70–86.

Chappelle, C.A., & Abraham, R.G. (1995). Cloze method: What difference does it make? In H.D. Brown & S. Gonzo (Eds.), *Readings on second language acquisition* (pp. 389–414). Englewood Cliffs, New Jersey: Prentice Hall Regents.

Graham, J.G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly, 21*, 505–521.

Hassan Basri Awang Mat Dahan. (2002). *Language testing: The construction and validation.* Kuala Lumpur: University of Malaya Press.

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

Kunnan, A.J. (2000). Fairness and justice for all. In A.J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). UCLES/Cambridge: Cambridge University Press.

Loh, E. F. (2006). *The effects of testing methods on the comprehension performance of ESL readers*. (Unpublished doctoral thesis). University of Malaya, Malaysia.

Oller, J.W. Jr. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal, 56*, 151–158.

Oller, J.W. Jr., & Jonz, J. (Eds.) (1994). *Cloze and coherence*. Lewisburg, PA: Associated University Presses.

Steinman, L. (2002). Considering the cloze. *Canadian Modern Language Review*, *59*(2), 291–301.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.

Yamashita, J. (2002). Mutual compensation between L1 reading ability & L2 language proficiency in L2 reading comprehension. *Journal of Research in Reading, 25*, 81–95.

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