Journal of English Language Teaching and Learning
Year 53 No. 217

Cloze validation against IELTS Reading Paper: Doubts on correlational validation *

Dr. Karim Sadeghi**

Abstract

Cloze was officially introduced in a journal on Journalism as a technique for estimating text readability and as "a new psychological tool for measuring the effectiveness of communication" (Taylor, 1953: 415). Different varieties of cloze have since been developed and experimented upon as measures of such diverse traits as reading comprehension and language proficiency. The findings of numerous corrleational studies on cloze as a measure of either skill is at best unsatisfactory and indeed contradictory. The present study seeks to find an answer to the question of whether standard cloze (with different text difficulty levels) is a valid measure of EFL reading comprehension (with IELTS Reading Paper as the criterion). 76 junior and senior students majoring in English Language and Literature at Urmia University participated in the study, where they sat 3 versions of standard 5-th deletion rate cloze tests as well as the Reading Paper of an Institutional IELTS (UCLES, 1995, 1997). While the results are in accordance with most previous research findings that cloze is a valid measure of EFL reading comprehension, serious problems are identified and discussed on the appropriacy of such a validation technique as correlation.

Key words: cloze test; correlation; validation; EFL learners; assessing reading

---

*- تاریخ وصول: ۱۳۸۹/۴/۲۸ تأیید نهایی: ۱۳۸۹/۵/۲۴

**- Assistant professor of Urmia University

1. ***Background.*** Cloze tests were initially developed more than half a century ago by Taylor (1953) to measure the readability of L1 texts. Soon after, the technique caught the attention of scholars in different fields and cloze procedure was accordingly applied for measurement purposes in L2. Cloze procedure was in particular used for gauging EFL learners' language proficiency as well as their reading ability. Much research was conducted on cloze as a measure of the either ability and the ultimate conclusion was summarised in the following statement statement: nobody still knows what cloze tests measure (Farhady, 1983; Lee, 1985; Sadeghi, 2008).

Different versions of cloze have been proposed and experimented by various researchers including standard cloze (also refereed to as fixed-ratio), rational cloze, discourse cloze, multiple-choice close, etc. All such cloze tests have been claimed to measure different traits mainly based on correlations between their results and those of other tests. While other verities of cloze still experience experimentation, the original every n-th deletion cloze (generally referred to as standard cloze) has widely been focused upon primarily because it retain the original concept of the term cloze itself. The fact that cloze tests are still being used to test reading comprehension in such widely-recognised tests warrants its further investigation. One such well-known instance is CPE (Certificate of Proficiency in English) constructed by UCLES, now known as Cambridge ESOL (UCLES, 2002), in which three four-choice cloze tests are used to test reading comprehension.

As far as cloze as a measure of reading comprehension is concerned, much criterion-related validation has led to contradictory findings, and it is yet to be known whether cloze tests can appropriately measure EFL reading comprehension or not. Among the scholars who have supported cloze as a valid measure of reading comprehension primarily because of its correlation with other supposedly valid tests of reading comprehension, one can name Greene (2001), Oller and Jonz (1994) and Davies (1979). The current study was, therefore, mainly an attempt to find out whether the

criterion-related validity of cloze as a measure of EFL reading comprehension is verified in an Iranian context, using a better recognised and more valid test of reading comprehension as a criterion measure, i.e. IELTS Reading Paper, against which not enough validation studies have been conducted. This study was accordingly an attempt to answer the following research question:

Is there any statistically significant relationship between standard cloze test and the IELTS Reading Paper as a measure of EFL reading comprehension?

This general research question was further divided into 3 more specific questions as follows:

1. Is there any statistically significant relationship between easy standard cloze test and the IELTS Reading Paper as a measure of EFL reading comprehension?

2. Is there any statistically significant relationship between medium standard cloze test and the IELTS Reading Paper as a measure of EFL reading comprehension?

3. Is there any statistically significant relationship between difficult standard cloze test and the IELTS Reading Paper as a measure of EFL reading comprehension?

The relevant null-hypotheses were tested at the probability level of 0.01.

2. Method

2.1 Participants. The participants were senior and junior EFL students majoring in English Language and Literature (ELL) at Urmia University, Iran. A total of 76 participants (41 seniors and 35 juniors) participated in the study. Table 1 below shows the characteristics of the subjects in terms of age, sex, first language background and other languages spoken.

Table 1: Characteristics of the participants

Characteristics / *Groups*	*Sex*		*Age*				*L1 background*		
	M	F	Mean	Range	SD	V	Azari	Farsi	Kurdish
juniors	16	19	24.5	12	4.32	18.7	19	6	10
seniors	12	29	24.57	7	1.74	3.03	25	5	11
Total	28	48	24.55	12	2.65	7	44	11	21

M: male; F: female; SD: standard deviation; V: variance; Other languages reported to be spoken: French (2), Turkish (16), Arabic (2).

All subjects spoke English in addition to the other language(s) they spoke.

2.2 Instruments. The main data elicitation measures (Mackey and Gass, 2005) used in the study were three standard cloze tests (called Sections A, B, and D) and the IELTS Reading Paper (UCLES, 1995, 1997) as the criterion reading comprehension test (Section C). An account of how cloze tests were constructed and a brief description of the criterion test used come next.

2.2.1 Cloze tests. To ensure that the selected texts were of enough interest to candidates and also of an authentic nature, cloze passages were chosen from texts intended for testing reading comprehension in one of the UCLES EFL examinations or Practice Tests (such as CAE and CPE), which use materials claimed to be authentic and of interest to general EFL candidates.

The difficulty levels of the texts used for making cloze tests and those in the criterion measure were calculated using the Smog Formula (Rye, 1982). As Rye (1982: 14) rightly emphasises, the readability of a text can be affected by many factors such as one's ability and desire to read, sentence length, word length, word frequency, subject matter, organisation of the material, syntax, physical environment, type of print, column size and line spacing, and angle at which the book is held. To these can be added the reader's familiarity with the writer's tone and style, his/her mental, emotional, and physical state at the time of reading, the time pressure under

which he/she is reading, and many others. Despite all such considerations, readability formulas are still widely used to measure the difficulty level of many texts.

Regarding the difficulty level of the texts used to make cloze tests, one text was comparable as much as possible in difficulty to the average readability of the texts used in the criterion measure (the IELTS Reading Paper); of the other two, one was more difficult and the other was easier than the average readability of criterion texts. The first cloze passage, with the title of 'Travel Companions', was adopted from UCLES (1996). The readability of this text was 8.65 in the Smog Index. The second cloze passage with the title of 'The urban revolution' was adopted from de Witt (1995). The readability of this text was 10.07 in the Smog Index. The last cloze passage, 'Sports writing', was taken from UCLES (2000b) with the readability of 11.

The above three passages were made into cloze tests using every 5th deletion rate. The first and the last sentences in each case were left undeleted to act as what are usually referred to as lead-in and lead-out, and the deletion began from the fifth word of the second sentence as is normal in standard cloze procedures. Care was taken not to delete numbers and proper names if they could not be inferred from the remaining context. In such cases, the word that followed was deleted. The relevant cloze tests appear in the appendix.

2.2.2 The criterion test. The validity of the experimental test in correlational research mainly depends on the validity of the criterion test against which the new test is validated. Although with the present state of our knowledge on the nature of reading comprehension, it is premature to claim that there is a valid and reliable test of EFL reading comprehension against which newly constructed tests can be validated, to choose the best criterion measure, a few well-known English proficiency tests (such as paper-based TOEFL, FCE, CAE, CPE, ELBA and IELTS) were reviewed and IELTS was found to be the best for our purposes. (The review of these tests has been reported in a paper which is under review in a different journal). What comes

next is a short description of the history, structure and content of IELTS.

The International English Language Testing System (IELTS) is the revised version of the English Language Testing Service (ELTS). ELTS, sponsored jointly by the British Council and UCLES, was originally developed to determine EFL candidates' English ability in order to enter higher education in Britain (Wier, 1987: 28). It consisted of two sections: General, which was intended to test reading, listening, and use of English; and Modular, which was expected to measure skills in reading, writing, listening and speaking in a particular subject area such as medicine, technology, social sciences, etc. The General section was to be taken by all applicants, but each applicant chose to take the related modular section (p. 29). ELTS was substituted by IELTS because of its inappropriate theoretical basis. It was based on Carroll's (1978) guidelines which lacked empirical evidence and Munby's (1978) model of needs analysis which was problematic for testing purposes (ibid.). The transition of ELTS to IELTS was also demanded because of the lack of reliability and validity information despite the observation that ELTS was more face-valid than 'almost all existing standardized tests designed for similar purposes' (p. 30). ELTS was finally replaced by IELTS in 1989.

IELTS was originally intended to measure the English proficiency of prospective non-native post-graduate students intending to study in English-speaking countries (UCLES, 2000a: 23). To accommodate 'important new developments in testing theory' and also as a response to demands from other groups of candidates (i.e., those wishing to pursue their studies at undergraduate level and also those seeking employment rather than study), and also to demands from receiving institutions, it was further updated in April 1995. The most recent development is that of the computer-based IELTS (CBIELTS) introduced in selected test centres in 2001 (ibid.). IELTS is jointly sponsored by UCLES, the British Council, and the IDP Education Australia: IELTS Australia (p. 1).

IELTS is composed of four sections or Modules: Listening, Reading, Writing and Speaking. The order of Modules is always as stated. While all candidates take the same Listening and Speaking Modules, there are Academic and General Training Modules for Reading and Writing. The Academic Module is taken by those who take the test for study purposes but the General Training Module is taken by those taking the test for work or employment purposes. The emphasis in the Academic Module is to assess a candidate's readiness in terms of English language ability to take an undergraduate or post-graduate course in a higher education institution where English in the medium of education. The focus of the General Training Module is to estimate one's 'basic survival skills in a broad social and educational context' (UCLES, 2000a: 4). In each case, the first three Modules are to be completed in the same day, but the Speaking Module can be taken on the same day or up to two days later based on the facilities in the centre (p. 5). In the CBIELTS, Listening and Reading Modules are computer-based and the candidate can choose to take the Writing Module on screen or on paper (ibid.).

IELTS final scores are reported in terms of 9 band scores, ranging from Band 1: non-user, for a person who 'essentially has no ability to use the language beyond possibly a few isolated words' to Band 9: expert-user, for a candidate who 'has fully operational command of the language; appropriate, accurate and fluent with complete understanding' (UCLES, 2000a: 1). An additional band score (Band 0) is also possible, which is not reported in test results and is for somebody who does not attempt the test (p. 1). There is a band score for each Module, and an overall band score for the whole test. While band scores for Listening and Reading are reported in 'whole and half' bands, those for Writing and Speaking are given in whole bands only (p. 20).

The Academic Reading Module of the IELTS uses a variety of test tasks to measure candidates' ability to cope with academic reading requirements in undergraduate and post-graduate courses. There are

three passages, selected from magazines, journals, newspapers, and books, with a total length of 2000 to 2750 words (UCLES, 2000a: 7). Texts are intended for non-specialist readers and are claimed to be of general interest (ibid.). Texts and tasks gradually become harder and at least one of them contains 'detailed logical argument' (ibid.). One text may also include non-linguistic elements such as graphs, or diagrams (ibid.). Technical terminology, if present, is explained below the texts. There are 40 questions in a variety of formats including multiple-choice, short-answer, sentence completion, notes, summary, diagram/flow-chart/table completion, choosing from a 'heading bank', identification of the writer's claims, yes/no or not given questions, classification, and matching lists or phrases (ibid.). The total time for the Reading Paper is 1 hour.

Regarding reliability and validity, until recently there were no publicly available statistics. The test-producing body, however, claimed that reliability was assured 'through the training, certification and continuous monitoring of examiners' for Speaking and Writing Modules (UCLES, 2000a: 19), and that 'The analysis for both Writing and Speaking shows a very consistent pattern across different test versions over time' (UCLES, 1999-2000: 6). In recent years, however, reliability information has been made public. For instance, the reliability of Listening Modules used in 1999 range from 0.87 to 0.91. Reliability figures for Academic Reading Modules in the same year ranged from 0.81 to 0.86, and for General Training Modules they ranged from 0.85 to 0.88 (UCLES, 1999-2000: 6). The higher reliability of Listening Modules compared to Reading Modules is because of the higher number of candidates who take Listening Modules (ibid.). Jones (2001: 4) reports a much higher reliability: 'IELTS is currently estimated to have reliability of 0.94, with a SEM of 0.36 of a band score.'

To ensure IELTS and other EFL tests' reliability, UCLES emphasises adhering to certain procedures. Test development in UCLES involves the following stages: perceived need for a new test, planning phase, design phase, development phase, operational phase,

and monitoring phase (Saville, 2001: 5). Each phase may also include several other activities, like the development phase, which 'involves validation activities, such as trialling and analysis, and the planning for implementation in the Operational Phase' (p. 6). Test construction begins by 'commissioning of materials for question papers' (UCLES, 20001a: 4; 2000a: 24). The next step is selecting, vetting, and editing of the material to be tested. The pre-test is then constructed, the materials are pre-tested, and the test items are analysed. Some of the analysed items are rejected, some are revised, edited, and pre-tested again, and others with desired item characteristics are stored in an Item or Materials Bank. Before the actual test papers are constructed, a process known as 'standards fixing' is applied during the construction process of the IELTS (UCLES, 2000a: 24). During this stage, the materials in the Item Bank are made into Trial Papers in the form of either a 60-minute Reading or a 30-minute Listening test. These Trial Papers are tried out on 'representative IELTS candidates and the results [are] analysed in order to allow accurate Band Score conversion tables to be constructed' (ibid.). This process is a necessary stage in order to 'ensure the equivalence of Listening and Reading versions [of the IELTS] and the reliability of the measurement of each paper' (ibid.).

Jones (2001: 4), however, expresses concern over the reduced reliability of the Cambridge EFL examinations as a result of the introduction of more authentic materials and task-based testing techniques instead of objective and discrete-point test items. As he rightly confesses, such a reduction in reliability is justified because of the improved test validity which is far more important than reliability. Comparing IELTS to other UCLES EFL examinations in terms of reliability, Jones (2001) points out that because IELTS is a 'non-certified testing system' and covers a 'wider range of the ability continuum' it tends to produce higher reliability indices (p. 3). Jones (2001: 2) warns test users about the interpretation of reliability indices: Reliability is not a characteristic of the test itself, but 'of an administration of the test' to a particular group of candidates. The

same test can produce highly reliable results with a large sample of 'widely-ranging ability' than with a smaller homogeneous group of equal ability.

In his introduction to the UCLES EFL examinations, Davies (1987: 20) notes that the 'established history' of these examinations 'gives users confidence in their validity and interpretation'. Much emphasis is placed on the judgemental rather than empirical validity in IELTS and other Cambridge EFL exams. Test validity is ensured by claiming that all the stages of test construction (commissioning, editing, pre-testing, analysis, banking of items, standards fixing, and question paper construction) are strictly observed 'throughout the writing and editing process, carried out simultaneously in Australia and Britain' (UCLES, 2000a: 24). There is also a Validation Group in UCLES conducting short-term and long-term validation research on IELTS to ensure that the test 'meets acceptable criteria in relation to quality and fairness' especially regarding validity, reliability, impact, and practicality (p. 14). Test validation, which is an ongoing process of test construction, is strongly emphasised in all UCLES EFL examinations; and as a part of this validation process, 'Checklists are completed based on the work carried out' to ensure that the tests meet the satisfactory standards of a valid and reliable test (Saville, 2001: 7).

Enright et al. (2000: 49) propose that using a variety of texts and integrated task types in tests of reading comprehension increases the construct validity of the test. Compared to other tests of the kind reviewed above, IELTS is the only one which has tried to incorporate such integrated tasks, which makes it more construct-valid than others. As regards external validity, a number of predictive validity studies has been done on IELTS which point out that it is a good predictor of academic success in English (UCLES, 2000a: 23). Jones (2001: 3) admires Cambridge examinations on the account that, in each revision, they give 'increasing attention to the communicative use of language, the better contextualisation of test items, and the authenticity of texts and tasks.'

Because for security reasons the actual IELTS test papers are not made available even for research purposes, due to confidentiality,

security, and ethical issues (Taylor, 2001: 12), it was decided to use an institutional version of IELTS (UCLES, 1995, 1997) which is a retired test. Such specimens are claimed to be as valid and reliable as secure IELTS. Linda Guymer (an IELTS official at Cambridge University) answered my email request for access to IELTS as follows:

We do not give out past papers as some of our versions will be used again. Our specimen materials are written in the same way as our live versions. The material used ... are authored by UCLES and would have been used as live material had they not been published.

In addition to the cloze tests and the criterion measure described above, the participants also received a cover letter in which the purpose of the study was explained. A short questionnaire preceded the answer sheet, in which students were asked about their sex, age, and linguistic background.

2.3. Procedure. Cloze tests Sections A and B were administered to senior EFL majors on one day. Section A took 30 minutes and section B took 45 minutes for the majority of the subjects to finish. One day later, Sections A and B were administered to juniors. For this latter group, the time taken to answer Sections A and B was 35 and 40 minutes respectively. Each group sat the criterion measure (Section C) and the other cloze test (Section D) exactly a week later than their first session. The time allocated for Section C was exactly 1 hour as suggested by the test producer (UCLES) for both groups and Section D took 35 minutes for both groups to finish.

2.3.1 Scoring procedure. Cloze tests were scored using both exact-word and acceptable- word procedure. In exact-word method, only the word used in the original text was counted correct. Although minor spelling mistakes were not penalised, changes to the tense of verbs, and to plurality and singularity of nouns were considered incorrect. As the literature suggests that acceptable-word scoring may be more suitable for EFL learners, all cloze tests were re-scored using this procedure to see whether there is any significant changes in the results. For this purpose, a group of educated native speakers (7 in number) were asked to answer cloze tests. Then the acceptability of

their answers were judged by two experts. Those words considered acceptable by both experts were, therefore, used as the acceptable criterion answers to re-score the EFL subjects' cloze answers. Section C (IELTS Reading Paper) was scored using guidelines given by test producers.

3. Results

3.1 Descriptive statistics of the data. Table 2 presents descriptive statistics for the tests used in the study. Statistics for cloze tests (Sections A, B, & D) have been calculated for both exact-word and acceptable-word scoring procedures. Since the mean scores of seniors and juniors were not significantly different on IELTS Reading Paper, all the computations are reported for the whole group (two groups together) and the final analysis and discussions will accordingly focus on all participants considered together.

Table 2: Descriptive statistics for the tests

Statistics / *Test & Group*	*Mean*	*Median*	*Range*	*Standard Deviation*
Section A				
Exact-scoring	7.24	6	13	3.53
Accept-scoring	9.88	9	15	4.28
TPS: 56				
Section B				
Exact-scoring	10.32	10	14	3.62
Accept-scoring	11.68	11	15	3.97
TPS: 59				
Section D				
Exact-scoring	17	15	25	6.9
Accept-scoring	19.48	19	28	8.19
TPS: 55				
Section C	18.08	17	22	6.18
TPS: 38				

Sections A, B, & D: cloze tests used; SD= Standard Deviation;
Section C: The criterion test (Reading paper of the IELTS); TPS= Total Possible Score

3.2 Reliability of the tests. One of the major characteristics of an instrument used in research is reliability. Reliability is considered a

necessary precondition for validity (Shohamy, 1983) and Farhady (1983: 257) points out that 'an unreliable test cannot be valid.'

Reliabilities of the tests used here were calculated using both K-R 21 formula and K-R 20 or Cronbach alpha. These formulas assume that items in the test are homogeneous (testing the same underlying trait) and independent. If all the tests are supposed to test the same thing, be it reading comprehension or something else, then the first condition is met. The problem of using these formulas and other inter-item consistency formulas with cloze tests is that items in cloze are not independent of each other, which is a breach of one of the basic assumptions underlying the use of such formulas. Although these formulas are widely used to arrive at cloze reliability indices, the above caution should be borne in mind in interpreting cloze reliabilities, as Farhady (1983) emphasises. To alleviate such a problem, the Guttman split-half procedure has been suggested (Bachman, 1985: 543; Jonz, 1991: 8) as an alternative. Brown (1983) believes that there is not much difference between these types of reliabilities in terms of practical consequences. For comparison purposes, however, the results of all types have been given below in Table 3.

Table 3: Reliability of the cloze tests (exact-method scoring) and the criterion test used in the study

Test / *Reliability method*	*Section A*	*Section B*	*Criterion (Section C)*	*Section D*
K-R 21	0.502	0.358	0.772	0.767
Cronbach's alpha	0.655	0.628	0.812	0.779
Split-half (Sperman-Brown)	0.658	0.764	0.860	0.911
Split-half (Guttman's Lambda 4)	0.649	0.718	0.857	0.898

A surface look at the reliability table shows that regardless of the fact that cloze tests used in the study were long enough (with more than 55 items each), and were expected to produce satisfactory

reliability indices, this did not happen (at least in two cases when computed through K-R 21). Except for split-half reliability where Section D's reliability was a little higher than the criterion test, in none of the other cases did the reliability of cloze tests exceed that of the criterion test, despite the fact that all cloze tests were longer than the criterion test, a feature that increases test reliability.

The reason for a relatively higher reliability for Section D may be that Section D turned out to be the easiest cloze test, contrary to the fact that it was assumed to be the medium-difficulty text based on the Smog Readability formula. The highest reliability score was, therefore, gained with the easiest cloze test because all the subjects had a chance to try all or most of the items. Despite the fact that Section A was regarded as the easiest text according to the readability formula, it turned out to be the most difficult cloze, in which many items remained either unanswered or were answered incorrectly. Therefore, the easy cloze test (Section D) produced a larger variance as Table 2 indicates, and this larger variance has contributed favourably to the amount of reliability. This amount of reliability for a 55-item test (when computed through alpha and K-R 21), compared with another shorter test (Section C with 38 items) with even higher reliability, does not seem to be satisfactory, however. Although the split-half reliability of the easy cloze test seems to be acceptable, none of the reliability indices of other cloze tests is high enough. The lower reliability indexes of hard and medium cloze tests (i.e., Sections A and B) are partly based on the fact that both cloze tests used were relatively difficult tests which produced less variance which in turn affected their reliability adversely. These low reliability figures of cloze tests imply that 'standard' cloze may not be a reliable enough test, and may consequently be an invalid one for measuring what it is supposed to measure, i.e., the EFL reading comprehension as measured by the IELTS Reading paper.

3.3 Relationship between variables. To find out how far scores on each cloze test correlated with the criterion measure, a correlation

analysis was run between each cloze and the criterion, using the Pearson-Product moment formula. To find out whether 'standard' cloze in general (called 'general' cloze here) correlates with the criterion measure used, another correlation was run between the average cloze scores on Sections A, B, and D and the criterion measure. From this point on, contrary to how they were categorised at the beginning of the study, and based on test results, Section A is considered the difficult cloze, Section B the medium cloze, and Section D the easy test. Table 4 indicates the degrees of correlation between cloze tests (scored using exact-word method) and the criterion test. Table 5 shows the same for acceptable-word scoring of cloze tests.

Table 4: Correlation coefficients between cloze tests (scored through exact-word method) and the criterion measure

Test	*Section A*	*Section B*	*Section D*	*General cloze*	*Section C*
Section A	1.000	0.795*	0.651*	0.854*	0.362
Section B		1.000	0.751*	0.910*	0.603*
Section D			1.000	0.934*	0.714*
General cloze				1.000	0.658*
Section C					1.000

p=0.01 df=74 r critical= 0.487

Table 5: Correlation coefficients between cloze tests (scored using acceptable-word method) and the criterion measure

Test	*Section A*	*Section B*	*Section D*	*General cloze*	*Section C*
Section A	1.000	0.828*	0.703*	0.886*	0.460
Section B		1.000	0.717*	0.887*	0.547*
Section D			1.000	0.923*	0.672*
General cloze				1.000	0.611*
Section C					1.000

p=0.01 df=74 r critical= 0.487

Table 6 shows correlation coefficients between cloze tests scored through exact-word and acceptable-word procedures. As the figures

indicate, the high and significant coefficients in each case indicate that, as far as the tests used in this study are concerned, there is not a significant difference between scoring procedures used. This finding is in contradiction with that of Oller (1973) and Alderson (1983) who found that for non-native speakers acceptable-scoring procedure was significantly superior to exact scoring.

Table 6: Correlation coefficients between scoring procedures (exact-word vs. acceptable- word) for the cloze tests used in the study

Test	*Section A (exact vs. acceptable)*	*Section B (exact vs. acceptable)*	*Section D (exact vs. acceptable)*	*General cloze (exact vs. acceptable)*
Correlation	0.942*	0.961*	0.981*	0.970*

p=0.01 df=74 r critical= 0.487

These high coefficients indicate that exact-scoring procedure can be safely used instead of the acceptable-scoring procedure without much difference in results. It should be noted that such a conclusion should be interpreted in the context of this study only. Further research with different number and type of participants, different texts used for cloze tests with different difficulty levels, and different acceptability criteria used for scoring, is needed to support this tentative conclusion.

To test for significance in differences in correlations between a cloze test scored by both exact-word and acceptable-word procedures and a criterion test, the Hotelling *t*-test is usually used (Glass & Hopkins, 1984; Guilford & Fruchter, 1987). Table 7 indicates whether differences between exact- and acceptable-scoring correlations of each cloze with the criterion test are significant or not. As such, none of the differences between correlation coefficients of cloze tests (scored both through exact-word and acceptable-word procedures) and the criterion measure was found to be significant.

Table 7: Hotelling t-test for comparing correlation coefficients between cloze tests (scored through exact and acceptable methods) and the criterion test

Test	*Section A (exact) & Section C vs. Section A (accept.) & Section C*	*Section B (exact) & Section C vs. Section B (accept.) & Section C*	*Section D (exact) & Section C vs. Section D (accept.) & Section C*	*General cloze (exact) & Section C vs. General cloze (Accept.) & Section C*
Correlations	0.362 vs. 0.460	0.603 vs. 0.547	0.714 vs. 0.672	0.658 vs. 0.611
Intercorrelation	Section A (exact vs. acceptable): 0.942	Section B (exact vs. acceptable): 0.961	Section D (exact vs. acceptable): 0.981	General cloze (exact vs. acceptable): 0.970
Hotelling t	1.5583	1.3339	1.4625	1.3832

p=0.01 df=73 t critical= 2.819 accept.: acceptable-scoring

3.4 Answers to research questions. Based on the evidence presented above, the research questions put forward at the beginning of this paper can now be answered and the related null-hypotheses tested.

The answer to the first question can be found in Tables 4 and 5. Table 4 shows the degree of relationship between the easy cloze test scored through exact-word procedure and the criterion test, and Table 5 shows the same for acceptable-scoring of cloze. Both Tables show that the easy cloze (Section D) is significantly correlated with the criterion test. The correlation coefficient between the easy test and the criterion measure in Table 4 (r=0.714) means that about 51% of the total variance produced by the IELTS Reading Paper is shared by the cloze test. Statistically speaking, the easy cloze test is relatively highly correlated with the criterion test and the null-hypothesis is thus rejected, meaning that there is a significant relationship between easy 'standard' cloze test (scored by exact-word method) and the Reading Paper of the IELTS. However, 51% of shared variance is not good enough a criterion for deciding that scores on the cloze can be used *instead of* those on the reading test to talk about subjects' reading comprehension. As far as acceptable scoring is concerned, the correlation between the easy cloze and the criterion test is also significant with the shared variance of 45%. Again the significant

degree of relationship between these two variables allows us to reject the null-hypothesis statistically.

To answer the second question, again data from Tables 4 and 5 can be used. In exact-word scoring of the cloze test, the correlation coefficient between the medium cloze (Section B) and the Reading Paper is 0.603, which is statistically significant, with the shared variance of only 36%. This coefficient is smaller for acceptable scoring of cloze test (r=0.547), but the relationship is still significant. Both these coefficients suggest that the second related null-hypothesis be rejected too, meaning that there is a statistically significant relationship between the medium 'standard' cloze test and the Reading Paper of IELTS as a test of EFL reading comprehension.

Contrary to the above cases, none of the relationships between difficult cloze test (Section A, scored by both exact- and acceptable-word methods) and the criterion test is significant, although there is a trend for such a significance when cloze is scored by acceptable-word procedure. The amount of shared variance in exact scoring of the cloze test is only 13% and in acceptable scoring, it is 21%. All this supports the related null-hypothesis, meaning that there is no statistically meaningful relationship between difficult 'standard' cloze test and the Reading Paper of the IELTS.

To answer the main question of the study, a general cloze score was worked out, as stated before, by averaging each subject's scores on easy, medium and difficult cloze tests. This 'general' cloze, calculated for both exact- and acceptable-scoring methods, was correlated with the criterion measure to find out the degree of go-togetherness. As Tables 4 and 5 represent, the resulting correlation coefficient is significant in both cases, meaning that there is a statistically significant relationship between 'standard' cloze in general and the Reading Paper of IELTS. The shared variance in either case, however, is not as meaningful as in the previous cases, i.e., 43% for acceptable-word scoring and 37% for exact-word scoring. The above findings provide the following answer to the main research question posed above: *There is a statistically significant relationship between*

'standard' cloze test and IELTS Reading Paper as a measure of EFL reading comprehension.

4. Discussion

The result of the correlational analyses reported above indicates that 'standard' cloze is statistically significantly related to the Reading Paper of IELTS as a test of EFL reading comprehension. Such an observation in language testing research has been interpreted as meaning that both instruments measure the same thing (for example, see Oller, 1973; Shohamy, 1983). The preceding statistical data would, therefore, mean that *standard cloze test is a valid measure of EFL reading comprehension*.

Such correlational validation has, therefore, led researchers and testers to suggest that one test (i.e., a cloze test) can safely *replace* another test (i.e., a criterion reading test). The recommendation of cloze as a valid substitute for other measures of EFL reading comprehension (and language proficiency) has also been partly based on the assumption that cloze tests are more economical and practical in constructing, administering, and scoring. However, it is argued here that concluding that one test can *substitute* another simply based on a high degree of correlation coefficient is not appropriate.

The use of correlation for validation purposes, in which it is concluded that one test can substitute another, cannot be sustained for at least four reasons. First things first, the concept of correlation conveys merely a sense of relationship between two variables (Glass & Hopkins, 1984; Goehring, 1981; Guilford & Fruchter, 1978; Garrett & Woodworth, 1958; Brown and Rodgers, 2002). The presence of a high degree of relationship between two variables does not mean that they are the same or replaceable; rather, based on information from one variable, one may be able to make predictions about the other. If the correlation between scores on two tests is +1.00, which is very improbable in practice, the scores on one test can then be predicted with perfect confidence if the scores on the other are known. Such a

perfect correlation, however, does not mean that two test measure the same thing and can conveniently replace each other.

Secondly, if one is allowed to substitute two tests simply because they correlate highly, let us see what might happen in the following case. There is little dispute that top and clever students tend to get better scores than average and slow students in many school subjects. It follows that a group of students' scores in those subjects will tend to correlate highly with one another, because in all of them top students always tend to be ranked first, then average students, and finally slow students. This means that a high correlation is expected between this group's, or any other group's by that means, scores on subjects like maths, English, history, geography, etc. Now if the correlation results are interpreted the way researchers in the field of language testing have done, one will be able to substitute the above tests with one another. Accordingly, one can give a history test to this group and then talk about how good they are in English or maths. This will eventually mean that because all tests are highly correlated and can substitute one another, one can give only one test during a course, e.g. an English test, and he/she will then be able to talk about students' ability in all other subjects. Such an assumption will not only be against all educational measurement principles and standards, it will also be senseless to any sensible person. Still, such a practice in language testing has prevailed for a long time, managing to escape the inspection of researchers.

The third reason why the use of correlation for validation purposes may be invalid is due to the contradictory results gained through the application of the technique in different contexts. In other words, the inconsistent findings in research where one test has correlated very highly with another in one context, but not so well in another context, may itself be an indication that whatever the relationship between these two tests, it is not shown clearly by correlation. This phenomenon gains more colour when the same tests correlated in similar contexts lead to different coefficients. As an example, let us suppose that we correlated maths scores of a group of students in one

school with their English scores and found a very high and significant coefficient of 0.98. Let us further assume that we checked another group of students' scores in maths and English in another school, and this time we found a correlation of -0.84. Now if a decision is to be made as to a criterion coefficient based on the first instance, how can it be accounted for by the second case?

The final reason why the validity of correlation for substitution purposes is under question is the vicious circle observed in validation of this kind. Namely, while sometimes the so-called valid tests like TOEFL or other ESL examinations are used as criterion measures on which dictation and cloze tests are validated, these latter tests become criterion measures themselves later on, against which other TOEFL and ESL tests are validated. Such a 'back-validation' process produces twice as many problems because, first of all, correlation is used for validation in the first instance, and secondly, tests such validated, despite their true validity not being established, become supposedly valid criteria, against which some other tests are to be validated still using the possibly improper technique of correlation.

5. Conclusion

Although correlational is a established and viable tool for validation purposes in educational measurement in general and language testing in particular, the interpretations based on significant correlations between tests have been faulty and misleading (Sadeghi, 2006). In light of the doubts explained above as to the use of correlation for validating language tests, all the results reported so far in literature on the validity of cloze as a measure of EFL reading comprehension (or language proficiency) based on correlational studies should be taken with a grain of salt. A more viable solution to the problem of whether results such attained are tenable or not may be arrived at by including qualitative investigation of the nature of the phenomenon in question as proposed by Babaii and Ansary (2001) and Sadeghi (2008). Researcher research has been offered as a promising alternative validating tool (Sadeghi, 2004).

References

Alderson, J. C. (1983). The cloze procedure and proficiency in English as a foreign language. In J. W. Oller (ed.), *Issues in language testing research* (pp. 205-212). Rowley, MA: Newbury House Publishers.

Babaii, E. &, Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System, 29,* 209-219.

Bachman, L. F. (1985). Performance on the cloze test with fixed ratio and rational deletions. *TESOL Quarterly*, 19, 335-356.

Brown, J. D. (1983). A closer look at cloze: validity and reliability. In J. W. Oller (ed.), *Issues in language testing research* (pp. 237-250). Rowley, MA: Newbury House Publishers.

Brwon, J. D., & Rodgers, T. S. (2002). *Doing second language research*. Oxford: Oxford University Press.

Davies, P. (1979). Cloze tests. In M. J. Raggett, C. Tutt, & P. Raggett (eds.), *Assessment and testing of reading: Problems and practices* (pp. 62-69). London: Ward Lock Educational.

Davies, A. (1987). Review of Certificate of Proficiency in English. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (eds.), *Review of English language proficiency tests* (pp. 20-21). Washington, DC: TESOL.

de Witt (1995*). How to prepare for IELTS*. Manchester: The British Council.

Enright, M. K., W. Grabe, K. Koda, P. Mosenthal, P. Mulcahy-Ernt, & M. Schedl (2000*). TOEFL 2000 reading framework: A working paper*. Princeton, NJ: ETS.

Farhady, H. (1983). New directions for ESL proficiency testing. In J. W. Oller (ed.)*, Issues in language testing research* (pp. 253-268). Rowley, MA: Newbury House Publishers.

Garrett, H. E. & R. S. Woodworth (1958). *Statistics in psychology and education* (5th ed.). London: Longman.

Glass, G. V. & K. D. Hopkins (1984). *Statistical methods in education and psychology (2nd ed.)*. Englewood Cliffs, New Jersey: Prentice-Hall.

Goehring, H. J., Jr. (1981). *Statistical methods in education*. Arlington, Virginia: Information Resources Press.

Guilford, J. P. & B. Fruchter (1978). *Fundamental statistics in psychology and education (6th ed.).* Tokyo: McGraw-Hill Kogakusha.

Greene, B. B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading,* 24, 82-98.

Jones, N. (2001). Reliability as one aspect of test quality. In UCLES, *Research Notes No. 4* (pp. 2-5). Cambridge: UCLES.

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

Lee, Y. P. (1985). Investigating the validity of the cloze score. In Y. P. Lee, A. C. Y. Y. Fok, R. Lado, & G. Low (eds.), *New directions in language testing* (pp. 137-147). Oxford: Pergamon Press.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Oller, J. W., & Jonz, J. (1994). Why cloze procedure? In J. W. Oller & J. Jonz (eds.), *Cloze and coherence* (pp. 1-18). London: Associated University Press.

Rye, J. (1982). *Cloze procedure and the teaching of reading*. London: Heinemann Educational Books.

Sadeghi, K. (2004). Researcher research: An alternative in language testing research. *The Reading Matrix: An International Online Journal, 4,* 85-95.

Sadeghi, K. (2006). *Rethinking correlational validation.* Paper presented at 3rd TELLSI Conference, Kermanshah, Razi University.

Sadeghi, K. (2008). Measuring reading comprehension: The judgemental validity of cloze procedure. *Iranian Journal of Applied Linguistics, 11,* 115-132

Saville, N. (2001). Test development and revision. In UCLES, *Research Notes No 4* (pp. 5-8). Cambridge: UCLES.

Shohamy, E. (1983). Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew. In J. W. Oller (ed.), *Issues in language testing research* (pp. 229-236). Rowley, MA: Newbury House Publishers.

Taylor, L. (2001). External requests for access to UCLES data/materials. In UCLES, *Research Notes No 4* (pp. 12-15). Cambridge: UCLES.

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

UCLES (1995, updated 1997). *IELTS Specimen Materials.* The British Council & IDP Education Australia.

UCLES (1996). *CAE Sample Papers*. Cambridge: UCLES.

UCLES (1999-2000). *IELTS Annual Review (1999/2000)*. Cambridge: UCLES, The British Council, IDP Education Australia.

UCLES (2000a). *IELTS handbook.* Cambridge: UCLES, The British Council, IDP Education Australia.

UCLES (2000b). *CPE: Revised CPE specifications and sample papers.* Cambridge: UCLES, ALTE (Association of Language Testers in Europe).

UCLES (2002). *CPE handbook and sample papers.* Cambridge: CUP.

Weir, C. J. (1987). Review of English Language Testing Service. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (eds.), *Review of English language proficiency tests* (pp. 28-31). Washington, DC: TESOL.