Measuring Reading Comprehension: The Judgmental Validity of Cloze Procedure

Karim Sadeghi
Urmia University

Abstract
Cloze tests have been widely used for measuring reading comprehension, readability and language proficiency. There is still much controversy on what it really is that cloze measures. The result of much correlational research is contradictory and very unsatisfactory. Thus, with a qualitative orientation, this study attempts to look at the judgmental validity of cloze as a test of reading comprehension. To this end, a group of 32 native and non-native speakers of English sat a standard cloze test. The participants were expected to complete most of the blanks correctly if cloze measured reading comprehension properly, because the text had been intended for undergraduates while cloze-takers were all either PhD students or members of academic staff with a PhD. Surprisingly, the results indicated that none of the participants reached the minimum native speaker performance criterion of 70%. Invited to reflect on what they thought they were doing when reading the blanked text, most cloze-takers felt that the text they read was a puzzle or a guessing game. Provided with the deleted words and asked to re-read the text, they confessed that cloze reading was very different from the second reading. Further findings and implications for future research are discussed in the paper.

Key Words: Cloze procedure, Reading comprehension, EFL/ESL testing, Validation, Correlational studies, Judgmental validity

Background

Cloze procedure is officially 57 years old today. Some researchers, however, assert that cloze is much older than what is generally believed. Carroll, Wilds and Carton (1959), for example, attribute cloze to a German psychologist called Ebbinghaus (1897) who deleted syllables to test 'the degree of fatigue in mental functions' of school children (p. 5). According to Kelly (1969), throughout the history of language teaching, cloze type tasks have been used for teaching and testing (Jonz, 1990, p. 91). Since its formal introduction in a journalistic periodical by Wilson Taylor (1953), cloze has been warmly received both as a research tool and as a testing and teaching device.

Although the technique was meant as "a new psychological tool for measuring the effectiveness of communication" (Taylor, 1953 p.415), it soon turned out to be a cure-all tool for measuring such various constructs as readability, intelligence, general language proficiency, general reading ability, specific reading comprehension, retention of learning, etc. Shortly after, cloze procedure gained such a strong foothold in language testing research that, according to Klein-Braley and Raatz (1984, p.135), "the inclusion of a close [sic.] in a study was viewed as a guarantee of quality". The outcome of much research carried out on cloze procedure as a measure of any of the above abilities is at best contradictory. One of the areas in which research on cloze has produced unhappy results is the area of reading comprehension. The fact that cloze tests are widely used for testing reading comprehension necessitates their further investigation. Gooskens and van Bezooijen (2006), Bertram (2006), Daztjerdi and Talebinezhad (2006), Zulu (2005), Spear-Swerling (2004), Hagtvet (2003) and Zervakis and Rubin (2002) are among the most recent studies experimenting with cloze tests for measuring comprehension or favouring them for such a purpose. Similarly, a well-known high-stakes test, i.e., CPE (Certificate of Proficiency in English) makes use of three four-choice cloze tests for the purpose of testing reading comprehension.

Despite the fact that many experiments have been conducted with cloze as a measure of a variety of skills and abilities, nobody knows what cloze tests measure (Farhady, 1983; Lee, 1985). Based mainly on correlational techniques whereby cloze test results have been correlated with results of other supposedly valid tests, cloze tests have been assumed to measure the same abilities as the criterion tests. Particularly related to cloze as a measure of reading comprehension are the following which have concluded that cloze is a valid measure of reading comprehension: Greene (2001), Sasaki (2000), Oller and Jonz (1994), Jonz and Oller (1994), Bachman (1985, 1982), and Davies (1979) to name a few.

By equating readability with comprehensibility, Taylor (1956, 1957) himself was the first to argue that cloze tests can measure readers' text comprehension. Oller and Jonz (1994) analysed a series of papers and argued that cloze tests are sensitive to inter-sentential high-level comprehension, and macro-structure processing. In the course of three experiments, McKenna and Layton (1990) compared cloze results (in normal and scrambled order) with another multiple-choice (m/c) reading comprehension test based on the same passage. Using fifth-grader native speakers as their participants, and correlating cloze results with those of the m/c reading comprehension test, they found natural-order cloze tests to be a measure of inter-sentential comprehension (p. 376).

Greene (2001) used a rational cloze test to measure macro-level comprehension and 'global coherence' of a theoretical discourse in the field of economics (p. 95). Noting the inability of standard cloze tests to measure macro-processing in discourse encountered by English as L1 university students, Greene used cloze tests in which items were meant to test 'recognition of cohesive devices and the ability to draw inferences from other sentences'. The results suggested that cloze tests used were valid measures of theoretical discourse comprehension simply because they had similar mean, dispersion, and frequency distribution to other True/False (T/F) comprehension questions.

There are, however, other studies which cast doubt on the appropriacy of cloze as a measure of high-level comprehension skills. For instance, Shanahan et al. (1982) compared 'natural', 'scrambled' and 'intruded' cloze tests in three experiments and concluded that using cloze tests for global comprehension was unreasonable (p. 251). Lado (1986), who used a questionnaire as well as quantitative analysis of the results of cloze tests administered to college-level native speakers of English, found that cloze-taking process was very different from normal reading process (p. 136). His 70% pass-mark criterion showed that all readers but one failed cloze tests when he used exact-word scoring method. Based on the participants' judgments and quantitative analysis, he considered cloze tests inappropriate measures of native language proficiency and comprehension. Porter (1976) seems to support Lado's (1986) position when he writes that he can get some meaning from a newspaper in Italian but he would be unable to predict a single word in that language. Based on this evidence he argues that "it would be misleading to suggest that the CP [cloze procedure] was measuring comprehension" (p. 152).

There are a few other researchers who have taken a middle position and believe that cloze tests can measure both low-level and high-level reading comprehension. Sasaki (2000), for example, writes that cloze tests, if rigorously designed, are capable of measuring micro-level and macro-level processing. Similarly, Porter (1976) argues that cloze tests can tap language comprehension at all levels.

Despite the fact that a lot of research has been carried out on cloze, the trait that cloze tests really measure still seems to remain a mystery. The contradictory findings obtained so far seem to be the result of the application of improper techniques in investigating the problem. In other words, in the majority of the cases in which cloze tests have been considered valid measures of reading comprehension (or other language abilities), such conclusions have been arrived at as a result of correlating cloze results with those of the criterion measures (the validity of which may well have been under question). Based on the obtained correlation

coefficients, cloze tests have been claimed either valid or invalid measures of reading or other abilities supposed to be tested by criterion tests. Such a validation procedure as a result of which cloze tests have been proposed to substitute the criterion tests has been practised by many well-known language testers and researchers including Oller (1973) and Shohamy (1983).

While correlation coefficient shows only some kind of relationship between two or more variables, it is quite strange that with such moderate coefficients as 0.54 in Jonz's (1976, p.261) study (which shows a variance overlap of about 30%), cloze tests have been regarded as valid substitutes for the criterion test (a 3-hour placement test in Jonz's case) in line with Oller's (1973) argument that "if the correlation is sufficiently high, it would be appropriate to substitute one type of test for another". The validity of correlational techniques for validation purposes in language testing as a result of which the tests being validated are suggested to substitute the criterion test is, however, under question (Sadeghi, 2006). Despite Brown's (1983, 1994) cautions that "relying almost exclusively on the statistical techniques developed for other fields" (1994, p.194) is unreasonable in testing language and that such validity indices "do little to explain how and actually what cloze is testing" (1983, p.238), this kind of criterion-validity has prevailed validation studies not only in the area of cloze testing but also in the discipline of language testing in general.

What is needed is, perhaps, to investigate the problem at hand using different kinds of tools, i.e., qualitative techniques, as some researchers have long expressed a desire for. The need for qualitative investigation has been voiced and practised in different forms by a good number of researchers including Babaii and Ansary (2001), Sasaki (2000), and Storey (1997) among the others. Markham (1988, p.48), for example, asserts that "purely quantitative techniques do not necessarily mirror the internal thought processes of the subjects." It is accordingly the purpose of the present study to use a qualitative tool, i.e., collective interview, to account for the judgmental validity of cloze procedure as a measure of

reading comprehension, i.e. to find out what cloze-takers think a cloze test measures.

Method
Participants
The participants in this study included 32 proficient readers all from CARE (Centre for Applied Research in Education) in the School of Education at UEA (University of East Anglia) in Norwich, UK. There were 4 native speakers of English, two of whom were lecturers and senior lecturers in Education and the other two were PhD students in Education. The remaining participants were all non-native speakers of English from Saudi Arabia, Iran, Greece, Egypt, Malaysia, and Taiwan, all of whom PhD students in Education.

Materials
A 5th-deletion random cloze test was constructed using a familiar text from Bryman (2001, p.354). The first and the last sentences were left intact to yield what Oller and Jonz (1994) call lead-in and lead-out, and deletions began with the fifth word of the second sentence. This procedure produced a number of 41 cloze items. (See the appendix.)

Procedure
The cloze test was administered to the above participants during a seminar meeting. The readers were initially given 10 minutes to read the passage and complete the blanks and the time was extended during the session so that most of them had finished reading the text. They were not told, however, that it was a test-type. Then they were asked to read out their answers one by one for each item.

The researcher then gave the exact words used in the original passage and invited the participants to reread the texts with blanks completed. An interview followed afterwards during which the candidates were asked the following questions:
1. What did you think you were doing when you were completing the blanks?

2. What do you think that the text with blanks was testing?
3. Do you think that the blanked test was a proper test for testing your comprehension of the original passage?
4. If somebody is going to test your comprehension of a text, would you like your reading comprehension to be tested in this way?
5. How do you compare your second reading (with the blanks completed) with your first reading (with blanks)?

Findings

There is ample evidence in the literature (Farhady and Keramati, 1996; Sasaki, 2000; for example) that scoring cloze procedure using exact-word method yields significantly high correlations with acceptable-scoring method. Both acceptable and exact-word scoring methods were used to score the cloze tests in the study, since the reader is expected to comprehend the writer's message rather than the exact words of the text. Table 1 shows descriptive statistics for both exact- and acceptable-word scoring procedures.

Table 1

Descriptive statistics for native and non-native speaker performance on the cloze test and the overall performance based on item-types (global vs. local)

Cloze test	Mean All items/ All subjects	Mean Natives	Mean Non-natives	Mean Local items	Mean Global items	SD
Exact scoring	37.12	53.02	24.39	42.47	13	7.34
Acceptable scoring	48.78	70.73	31.22	51.65	40.67	10.05

SD= Standard Deviation
Note: For ease of comparison, all mean scores in the table are shown in percentage.

Based on the simple descriptive statistics above, it is clear that the performance of all the participants (both natives and non-natives) on the cloze test was far below the 70% pass mark criterion when the tests were scored using exact-word method. With acceptable-word scoring

procedure, however, only two native speakers were able to score just above the cutting-point. Their performance becomes much more meaningful when we recall that the text used as cloze was intended for undergraduate students.

The weak performance of the candidates on the cloze test in both scoring methods can be a sign of two things: either they were unable to comprehend the text on which the cloze test was based, or the cloze test was unable to measure their comprehension of the original text. To test which one of these assumptions could be verified, the deleted words were given to the participants and they were invited to re-read the text with the blanks completed. Some of the readers expressed their shock at how differently they had comprehended parts of the text when they first read it with blanks. None of them had any difficulty in understanding the text in this latter reading. They also noted that they spent much less time on the text when blanks were known compared to their first reading.

Reacting to the questions asked during the collective interview phase, the readers were more or less unanimous in that they considered the blanked text a puzzle to be solved or a kind of language game. Asked about what it was that they thought the blanks were *testing*, one of the native speakers expressed that he did not consider it a *test* at all. Some of the participants regarded it as a test of vocabulary; others considered it a test of grammar; and particularly non-native speakers felt that it was a test of English *language* because native-speakers, they argued, could provide more appropriate answers to the blanks mainly because of their higher *language ability*. A few other candidates felt that it was a tool for guessing the writer's style.

When asked whether they would be happy if their comprehension of a passage was to be tested using such a technique, all of the participants gave negative answers, preferring comprehension questions or writing a summary of the text. In general, when they compared their first reading (with blanks) with their second reading (with the blanks completed), most candidates felt that being forced to focus on the blanks and being

pushed to neglect the whole passage and the overall picture blocked their comprehension of the text. The following Table summaries the answers given by the respondents to questions asked during the interview.

Table 2
Answers given by cloze-takers to interview questions

Question	Candidate responses
Q. 1	Completing a puzzle; doing a language game
Q. 2	It was not a test at all; it tested the range of my vocabulary; it was a test of grammar; it was a test of English language ability; it was a way of guessing the writer's style
Q. 3	No
Q. 4	No; I prefer comprehension questions to be used instead; summary writing after reading the text is a better measure of reading comprehension than the blanked text
Q. 5	The second reading was totally different from the first reading; I found that my comprehension of the blanked test was very different when I read the text with no blanks; focusing on blanks prevents the reader from grasping the overall meaning

Discussion
Considering the fact that all the candidates in the study were reasonably expected to have been more proficient in reading comprehension than or at least as proficient as undergraduates (for whom the original cloze text had been intended), and the fact that the chosen passage was from a qualitative research methods book with the topic of which all the candidates were quite familiar, it is highly likely that all the readers would comprehend the original text without difficulty. It was

hypothesised that if cloze procedure could measure reading comprehension properly, all the candidates without exception should be able to complete all the blanks correctly (at least in acceptable scoring) without any challenge. A surface look at Table 1 above shows that this was far form being the case. Based on their cloze scores, neither native speakers nor non-natives seem to have understood the passage properly. This observation means that either the readers were not proficient enough to understand the passage or cloze scores did not reflect their comprehension properly, that is, cloze is not an appropriate measure of reading comprehension.

As noted earlier, the original text on which the cloze was based had been intended for undergraduates receiving training in qualitative research methods. The fact that the participants in this study were either research post-graduates or native speaker university lecturers leaves no doubt that their reading proficiency was enough for comprehending the text – and indeed, when the items were given to the candidates and they read the blankless text, it was revealed that the original text did not produce any challenge for any of the readers. As such, the first justification that the readers were not proficient enough to complete the cloze passage properly is out of place. The second explanation that cloze scores are not valid indicators of reading ability gains strength here especially because the qualitative findings in the interview phase indicated that none of participants felt that such a tool was appropriate for measuring their text comprehension.

Our findings here that standard cloze procedure is not an appropriate tool for measuring text comprehension runs counter to most other studies of cloze where cloze has been found or argued to be a valid test of reading comprehension. Most of the studies dealing with cloze as a measure of reading comprehension have regarded cloze as a valid measure of reading comprehension either because they have found high correlations between cloze results and other measures of reading comprehension (the validity of which may well be under question) or because they have used scrambled cloze and compared results with

normal-order cloze. The findings of the former group should be taken with a grain of slat because of the problem already discussed with the use of correlation for validation purposes. The problem with the latter type of studies is that they have equated inter-sentential connectivity and/or the ability to reconstruct cohesive ties with reading comprehension. Data from our study show that success in completing items which require inter-sentential integration (like items 18 and 21) may not require high-level comprehension; and conversely, the ability to restore items dependent on local context (such as items 6 and 28) can be far more demanding and challenging in some cases. Furthermore, despite having read the original text several times, and having no problem in its comprehension, the researcher himself could score only at 75.61% in exact-scoring and at 85.37% in acceptable scoring. Moreover, neither in exact nor in acceptable scoring was he able to fill in all so-called 'local' items correctly. All this suggests that whatever it was that the cloze test was measuring was something different from overall reading comprehension, and that the 'local' items did not test local comprehension only. Therefore, concluding that a cloze test measures reading comprehension or not simply based on the number of inter-sentential (local) or intra-sentential (global) items does not seem to be justifiable.

Another argument about the validity of cloze as a measure of reading comprehension is that cloze can be regarded as a proper and valid test of comprehension of a text only if cloze-taking involves the same processes as normal reading. This research presents evidence in that readers were acting differently when they were taking the cloze test compared to the time when they read the same text as a normal passage. A neglected point in research on cloze as a measure of reading comprehension is that while other testing methods do a fair job of presenting the complete text to the reader first and then try to find out if the text has been comprehended by the reader, cloze tests appear to be too unfair and unethical in this regard in that they require the reader to first reconstruct something hidden from him/her, and then to understand the rightly or wrongly reconstructed discourse. The results of the present study

suggest that cloze tests may be used to test various aspects of language, but using them for testing comprehension of a text that the readers have not yet seen in its undeleted form is unfair, invalid and unjustifiable.

It should finally be stressed that quantitative findings and scores by themselves do not reveal the true picture of the reality, and it would be nearly impossible to claim whether cloze or any other test can properly measure a certain trait by looking at scores or correlation indices and neglecting the mental processes happening at the moment of taking the test. 'Researcher-research' has been proposed as an alternative validation technique, which when combined with other qualitative procedures may provide the researcher with better insight on what it is that cloze is measuring (or not measuring) (Sadeghi, 2004).

Conclusion and Implications
While a need for further research is called for, the findings here show that cloze procedure has failed the judgmental validity test, which has far-reaching implications for researchers, testers, and teachers. All of these groups are recommended to consider critically their previous conceptions of cloze as a measure of reading comprehension. It should be borne in mind that economicality of a testing tool should be a desirable asset as far as its use for what it can properly measure is not under question. Therefore, favouring cloze tests over other techniques for measuring text comprehension is not tenable simply because cloze does not seem to measure *just* overall understanding.

References

Babaii, E. & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle?, *System*, 29 (2), 209-219.

Bachman, L. F. (1982). The trait structure of cloze test scores, *TESOL Quarterly*, 16 (1), 61-70.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletion, *TESOL Quarterly,* 19 (4), 535-556.

Bertram, C. (2006). Exploring teachers' reading competencies: A South African case study, *Open Learning*, 21 (1), 5-18.

Brown, J. D. (1994). A closer look at cloze validity. In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 189-196). London: Associated University Press.

Bryman, A. (2001). *Social research methods.* Oxford: Oxford University Press.

Carroll, J. B., Carton, A. S. & Wilds, C. (1959). *An investigation of 'cloze' items in the measurement of achievement in foreign languages*. Cambridge, MA: Graduate School of Education, Harvard University, Laboratory for Research in Instruction.

Dastjerdi, H. V. & Talebinezhad, M. R. (2006). Chain preserving deletion procedure in cloze: A discoursal perspective, *Language Testing*, 23 (1), 58-72.

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.

Farhady, H. (1983). New directions for ESL proficiency testing. In J. W. Oller (Eds.), *Issues in language testing research* (pp. 253-269). Rowley, MA: Newbury House Publishers.

Farhady, H. & Keramati, M. N. (1996). A text-driven method for the deletion procedure in cloze passages, *Language Testing*, 13 (2), 191-207.

Gooskens, C. & van Bezooijen, R. (2006). Mutual comprehensibility of written Afrikaans and Dutch: Symmetrical or asymmetrical?, *Literary and Linguistics Computing*, 21 (4), 543-557.

Greene, B. B. (2001). Testing reading comprehension of theoretical discourse with cloze, *Journal of Research in Reading*, 24 (1), 82-98.

Hagtvet, B. E. (2003). Listening comprehension and reading comprehension in poor decoders: Evidence for the importance of syntactic and semantic skills used for oral comprehension, *Reading & Writing*, 16 (6), 505-539.

Jonz, J. (1976). Improving on the basic egg: The M-C cloze, *Language Learning*, 26 (2), 255-265.

Jonz, J. (1990). An intratextual analysis of cloze procedure, *Southwest Journal of Linguistics*, 9 (2), 91-106.

Jonz, J. & Oller, J. W. (1994). A critical appraisal of related cloze research. In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 371-407). London: Associated University Press.

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

Lado, R. (1986). Analysis of native-speaker performance on cloze tests, *Language Testing*, 3 (2), 130-146.

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.

McKenna, M. C. & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension, *Journal of Educational Psychology*, 82 (2), 372-377.

Markham, P. L. (1988). The cloze procedure and intersentential comprehension in college-level German, *IRAL*, 26 (1), 44-51.

Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure, *Language Learning*, 23 (1), 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.

Porter, D. (1976). Modified cloze procedure: A more valid reading comprehension test, *ELT Journal*, 30 (2), 151-155.

Porter, D. (1978). Cloze procedure and equivalence, *Language Learning*, 28 (2), 333-341.

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

Sadeghi, K. (2004). Researcher research: An alternative in language testing research, *The Reading Matrix: An international Online Journal*, 4 (2), 85-95.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple-data source approach, *Language Testing*, 17 (1), 85-114.

Shanahan, T., Kamil, M. L. & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension, *Reading Research Quarterly*, 17 (2), 229-255.

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

Spear-Swerling, L. (2004). Fourth graders' performance on a state-mandated assessment involving two different measures of reading comprehension, *Reading Psychology*, 25 (2), 121-148.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test, *Language Testing*, 14 (2), 214-231.

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

Taylor, W. L. (1956). Recent developments in the use of 'cloze procedure', *Journalism Quarterly*, 33 (1), 42-49, 99.

Taylor, W. L. (1957). 'Cloze' readability scores as indices of individual differences in comprehension and aptitude, *Journal of Applied Psychology*, 41 (1), 19-26.

Zervakis, J. & Rubin, D. C. (2002). Production and recognition of bias of stylistic sentences using a story reading task, *Journal of Psycholinguistic Research*, 31 (2), 107-130.

Zulu, C. (2005). Academic reading ability of first-year students: What's high school performance or prior exposure to academic reading got to do with it?, *Southern African Linguistics & Applied Language*, 23 (1), 111-123.

Appendix

Language in qualitative research

Language is bound to be of importance for social researchers. It is after all (1)through language that we ask (2)people questions in interviews and (3)through which the questions are (4)answered. Understanding language categories has (5)been an important component of (6)research involving participant observation, because (7)knowing how words are used (8)and the meanings of specific (9)terms in the local vernacular ((10)often called 'argot') is frequently (11)viewed as crucial to an (12)appreciation of how the social (13)world being studied is viewed (14)by its members.

In this (15)chapter, however, two approaches will (16)be examined that treat language (17)as their central focal points. (18)They are called conversation analysis (19)(CA) and discourse analysis (DA). (20)What is crucial about these (21)approaches is that, unlike traditional (22)views of the role of (23)language in social research, they (24)treat language as a topic (25)rather than as a resource ((26)admittedly a clichéd phrase). This (27)means that language is treated (28)as significantly more than as (29)a medium through which the (30)business of social research is (31)conducted (such as asking questions (32)in interviews). It becomes a (33)focus of attention in its (34)own right. While CA (35)and DA do not exhaust (36)the range of possibilities for (37)studying language as a topic, (38)they do represent two of (39)the most prominent approaches. Each (40)has evolved a technical vocabulary (41)and set of techniques. This chapter will outline some of the basic elements of each of them and draw attention to some contrasting features.