한국산학기술학회논문지
Vol. 12, No. 2 pp. 640-645, 2011

Making Cloze Tests More Valid

Chung-yeol Park[1*]
[1]**Owens International College, Korea Nazarene University**

클로즈 테스트의 효율성 높이기

박정열[1*]
[1]나사렛대학교 오웬스국제대학

Abstract The purpose of this article is to investigate ways of making the cloze test more valid as an English proficiency test so that it can provide a better format for the cloze test especially to the non-native teachers in EFL environments. This investigation identifies ways of validity in terms of materials, deletion rates, item characteristics and scoring methods used in cloze tests. An increase in test validity can be accomplished by knowing how the four factors affect the cloze test results and how to manipulate them. Practical suggestions are provided for teachers to design a cloze test. Teachers can develop the reliability and validity of cloze tests by manipulating the four factors. In seeking a more valid and reliable cloze format, this research gives some suggestions based on a literature review. We can conclude that a cloze test can be valid only when teachers are aware of the factors which affect the test result and the ways to handle the factors to make the test appropriate for their test purposes.

요 약 본 논문의 목적은 영어 구술시험에서 클로즈 테스트가 더욱 효과적으로 활성화될 수 있는 방법을 조사하여 영어를 외국어로 학습되는 환경에서 영어가 모국어가 아닌 교사들이 클로즈 테스트를 더욱 활용할 수 있게 하기 위함이다. 이는 클로즈 테스트에서 사용되는 자료, 삭제율, 항목의 특징, 그리도 점수 계산 방식을 조사하는 것이다. 테스트의 효율성을 향상시키는 것은 어떻게 위의 네 가지 요소가 클로즈 테스트의 결과에 영향을 끼치는 지를 파악함으로써, 그리고 그것들을 어떻게 잘 조절할 수 있음으로써 가능할 수 있다. 본 논문은 교사들이 네 가지 요소들을 능숙하게 다루어 클로즈 테스트의 효율성과 신뢰성을 발전시킬 수 있도록 몇 가지 제안을 하였다. 결국 중요한 것은 교사들이 테스트의 결과에 영향을 주는 요소들을 잘 알고, 그들의 목표에 알맞은 테스트를 만들어 내기 위해 네 가지 요소들을 잘 조절함으로써, 클로즈 테스트의 효율성과 신뢰성을 높일 수 있다는 것이다.

Key Words : English Proficiency Test, Reliability, Validity, EFL Environment

1. Introduction

Though cloze tests are widely used because they are easy to construct, administer and score[10], many cloze tests have resulted in a failure to measure students' English proficiency due to the lack of knowledge and the misunderstanding of the test in Korea. Although many studies have proved that a cloze test is an effective means of measuring proficiency[3,5,7,9,12,14], not all cloze tests can be automatically assumed to measure proficiency. As Brown points out, natural cloze tests, which are defined as cloze procedures developed without intercession based on the test developer's knowledge and intuitions about passage difficulty, are not able to measure what the cloze test intends to[5]. This claim is also supported by Alderson's study which examined the effect of three

This work was supported by Korea Nazarene University
*Corresponding Author : Park, Chung-yeol(cyp4x4@kornu.ac.kr)
Received November 18, 2010 Revised January 24, 2011 Accepted February 10, 2011

variables of cloze tests: material difficulty, deletion rate, scoring method on the test validity[2]. From this research, it was apparent that the correlation coefficient between the cloze test and the proficiency test which he used as an external criterion was highest with the easy text, 12 deletions and semantically acceptable score method. However, the correlations were significantly varied as the variables were manipulated. Therefore, it can be concluded that extra care is necessary when constructing and interpreting a cloze test because the results may have been affected by the particular set of words deleted, the scoring method, and the passage used[11].

However, many classroom teachers have difficulties in choosing the passage for a cloze test. In addition, the decision of selecting the words which would be deleted and the way of deletion have also been a challenge to the teachers. The most difficult aspect for them in the cloze test has been the scoring. Though the teachers know the correct answer, they may have been overwhelmed by the variety of close answers with which many students used in the blanks. The classroom teachers may have struggled to distinguish between the correct and incorrect answers. These difficulties can be attributed to the lack of knowledge about the correlation between the test procedures and the test results.

The purpose of this article is to investigate ways of making the cloze test more valid as an English proficiency test so that it can provide a better format for the cloze test especially to the non-native teachers in EFL (English as a foreign language) environments. This investigation will identify ways in terms of materials, deletion rates, item characteristics and scoring methods used in cloze tests. Material will be dealt with from the perspective of its content and level of difficulty. An increase in test validity can be accomplished by knowing how the four factors affect the cloze test results and how to manipulate them.

Most studies on cloze tests have covered the four factors separately with different results, and classroom teachers have been confused and cannot reach their own conclusions. What is relatively unique about this study is that it focuses on the practical usage for classroom teachers who have been attracted to cloze tests due to large classrooms and a limited time. This article will present desirable cloze test procedures based on a comprehensive literature review.

2. Body

The first thing that teachers have to do to design a cloze test is to select the material. The following research gives them criteria for selecting cloze material. The findings of Sasaki demonstrate the importance of cultural schemata in students' test-taking processes[13]. The results show that those who read culturally familiar cloze texts tried to solve more items whether the answers were correct or not and generally understood the text better, which resulted in better performances than those of the students who read the original text which was not culturally modified. The familiar group utilized within-sentence information to obtain correct answers significantly more often than the unfamiliar group. The familiar group also used other more extensive information more frequently than the unfamiliar group. According to Yuet, the cultural content is more critical for beginners or intermediate level readers than for high level readers[15]. It seems that readers with lower proficiency rely more on background knowledge.

Chihara also found that simple things like nouns referring to persons and places carry more subtle semantic and pragmatic information[7]. They changed several culturally unfamiliar terms (Joe, Athen, Klein's) from two English cloze tests into more familiar terms (Hiroshi, Osaka, Daiei) for Japanese participants. The participants significantly improved their performance on the modified cloze test than on the original tests across two different types of content.

These results demonstrate how non-native English teachers in an EFL environment should select a cloze material. Non-native English teachers may not be able to write passages for themselves out of fear of making mistakes and lack of authenticity. Authentic materials in books or on the Internet could be the best source for them. However, the authentic materials in many cases inevitably should have cultural content that may affect the cloze scores. Therefore, modification is needed for a cloze test especially for lower level students. Modifications should neutralize the effect of cultural familiarity so that the test results can represent students' pure proficiency.

Pronouns can be changed into the names of the native language and the passages that contain specific customs should be avoided. English newspaper articles which introduce students' native cultures or news are a valuable source for advanced students.

The difficulty of the text has to be appropriate to students' levels. Teachers should not assume that all cloze passages regardless of difficulty can measure the students' proficiency. Alderson shows teachers the importance of the issue by identifying the correlation between text difficulty and the cloze test[2]. The correlation coefficient between the material difficulty and the ELBA scores as an external criterion for a proficiency test was changed as the material difficulties varied. Moreover, the results show that the cloze test can measure the different components of language proficiency according to the text difficulty. For example, the difficult text correlated consistently higher with the ELBA tests especially in grammar, vocabulary and reading comprehension compared with medium and easy text. The findings imply that teachers have to be careful in deciding on the level of text difficulty so that it can fit into the students' proficiency level and the purpose of the cloze test.

The above research also sheds light on the deletion rate. The correlation coefficient between the cloze test and the dictation test which was used as an external criterion was changed by varying the deletion rate. The correlation coefficient was changed from .40 with a deletion rate of 8 to .91 with a deletion rate of 12, which means changing the deletion rate can have a significant effect on the validity of a cloze test.

Teachers also have to know the characteristics of fixed ratio and rational deletions. Bachman developed a classification framework for cloze item types according to the hierarchical context hypothesized as necessary to complete each item: 1) within clause, 2) across clause, within sentence, 3) across sentence, within text, 4) extra-textual[4]. He found that the difficulty level of item types accorded with the hypothesized order of difficulty according to the level of context required, i.e., type 1) being the easiest and the type 4) being the most difficult in a rational deletion cloze test. In addition, a fixed-ratio cloze passage tends to contain more items that can be filled in simply by using clause-level grammatical knowledge or extra-textual knowledge types 1) and 4) than items that require the ability to use types 2) and 3). It seems to be reasonable for teachers to measure global comprehension ability to via rational deletions rather than a fixed-ratio procedure.

Abraham and Chapelle also support the above results[1]. They examined the amount of context that is required to restore the word with different cloze formats: fixed-ratio, rational deletion and multiple-choice cloze. According to the conclusion of the study, the meaning of the fixed-ratio cloze scores can be thought of as the students' ability to retrieve content words from long-term memory or to find them elsewhere in the text. The scores also indicated the ability to produce words in their correct morphological form. In contrast, performance on the rational cloze items was affected by the context levels of their clues. Thus, the scores can be interpreted as indicating ability to identify contextual clues.

We can conclude that rational-deletion is a better deletion procedure than the fixed-deletion because it can utilize context level clues. The fixed deletion rate should be avoided for the valid cloze test result. When teachers use the fixed deletion rate, the items may not measure the students' proficiency as a whole, which means the test could result in a discrete- point test which can only measure fragmental grammar knowledge about English.

The other thing that teachers have to be mindful of regarding deletion is that they have to be careful in deciding on selecting the first-deleted word. The first cloze item is very important because it can provide the first clue for test difficulty for students. A difficult item has to be avoided as a first item for students not to be excessively nervous or expect failure. To provide the context or activate background knowledge which is needed to restore items, several sentences for the first and last part should be left intact.

A cloze test can function as a discrete-point test or diagnostic test when teachers design it appropriately for their purpose. For example, teachers can delete words that focus on specific parts of speech such as prepositions, articles or content words, which they would like to strengthen or test students' vocabulary on a specific topic according to students' level and/or their strong and weak points.

The above research emphasizes that teachers should account for the reasons why they delete the words when

they develop a cloze test. This means that they need to know the item characteristics that affect the test results. Some researchers investigated the relationships between item characteristics and the cloze test result.

Teachers have to be aware of what kinds of words are involved in a particular cloze test so that they can understand and predict what the test results can tell them. The first significant characteristic that teachers should consider when they select words-deleted is whether the words are content or function words. According to Kobayashi, the content words were difficult to be restored[11]. The result is also supported by the study of Abraham and Chapelle[1]. They found that function words were easier than content words both in the fixed ratio and the rational deletion cloze test. Therefore, when the purpose of a cloze test is to measure student global proficiency rather than specific grammar or the knowledge of English, content word items which require more knowledge and context to be restored should be deleted rather than function words.

Kobayashi also reported that parts of speech are an important factor that can significantly affect the cloze test results[11]. Difficult items were the relative pronouns, pronouns, and articles. Articles were the most difficult item. She attributed the cause to the native language of the subjects, i.e., Japanese, which does not have any articles. Therefore, when teachers consider the item difficulty, it is recommended that they contrast English with their native language.

Easy items were nouns and verbs when syntactic variations like verb endings, tenses and plural forms were permitted. Item-total correlations became high on the item of to-infinitives, adverbs and conjunctions because they are related to textual organization requiring a high level of language ability. Categorizing the items into parts of speech can be a useful way to adjust a cloze test difficulty and predict the test results when teachers design a cloze test.

Brown examined cloze test difficulties related to word length in terms of the number of letters[6]. The length of words was correlated negatively with item difficulty. Abraham and Chapelle reached the same conclusion especially in the rational deletion cloze test[1].

According to Kobayashi and Abraham, the items which allowed alternative answers were more difficult than those that did not because they need more cognitive ability[1,11]. Regarding the number of occurrences and frequencies, the above studies which were conducted by Brown, Abraham and Chapelle, Kobayashi yielded the same result; the occurrences and frequencies were inversely proportional to the level of item difficulty[1, 6, 11].

The amount of context required to restore the word is the another important item characteristic which can affect the test score. The amount of context was correlated negatively with item difficulty in the rational deletion cloze test[11]. This finding suggests that items tend to be easier when a smaller amount of context is necessary to restore the words.

Content word was difficult on the item which requires semantic knowledge while function word was easy on the item which requires syntactic knowledge. Hence, it is needed for teachers to subcategorize the content and function word based on the kinds of knowledge which are needed to restore the item.

As we have seen above, the relationship between the item difficulty and item characteristics is an important subject that teachers have to consider. As Kobayashi reported[11], article items, which were most difficult for Japanese test takers, did not yield significant item discrimination. When a cloze test is only filled with difficult items which do not have item discrimination, it results in a meaningless test for teachers as well as students. Therefore, it is recommended that teachers pay attention to students' language performance and errors in the classroom in order to adjust item difficulty and discrimination.

One of the strong points of the cloze test is that it is easy to score. However, teachers must not be tempted only by the ease of scoring without considering fair and reasonable scoring methods. Kobayashi shows that the acceptable-word scoring methods led to higher mean scores[11]. She applied three different scoring methods; exact-word scoring method, semantically-and-syntactically -acceptable-word scoring method, and semantically -acceptable-but-syntactically-unacceptable-word scoring method. The semantically-acceptable-word scoring method yielded consistently higher scores. The two acceptable word scoring methods had consistently higher reliability, although the differences were not very great. Content

word items had lower reliability when the exact-word scoring method was applied. However, values rose dramatically when acceptable-word scoring methods were applied. Furthermore, the result of the comparison across proficiency levels showed that higher groups gained more in the acceptable-word scoring methods. The gains from the exact-word scoring method to the acceptable-word methods were greater for higher groups. Middle groups benefited from both of the acceptable-word methods. Another finding is that lower proficiency groups benefited more than the higher groups when the semantically -only-acceptable word scoring method was applied. This suggests that lower-proficiency learners may have had more problems at a syntactic level than higher-proficiency learners. These findings support the view that the acceptable-word scoring method is fairer especially to those with a greater language proficiency.

Oller also supported the findings[12]. Correlation coefficients between cloze test and dictation test as an integrative and proficiency test were highest when the contextually-acceptable method was applied. Additionally, Stubbs and Tucker suggest that non-native English teachers can use the contextually-acceptable method by showing the high correlation coefficient between the scores of exact-word only method and contextually-acceptable method[14]. Therefore, cloze tests seem to become more valid as a measure of reading comprehension based on the above research when emphasis is on meaning rather than linguistic accuracy.

However, teachers should not easily conclude that acceptable scoring method is always preferable. Students tend to ignore the accuracy and/or appropriateness when they are not rewarded or penalized by scores. Teachers have to consider the backwash effect when they apply a scoring method. When students do not have proper accuracy or they have fossilized grammar errors, the exact-word scoring method or semantically-and-syntactically-acceptable scoring method could help students to correct syntactic errors.

Extra care is needed for valid scores when non-native teachers apply the acceptable score method and would like to give partial credit. Initial-letter cue format which can limit the possible answers can be a way not to be overwhelmed by the close and confusing answers for non-native English teachers.

3. Discussion

In this paper, I have investigated ways of making a cloze test more valid in terms of material, deletion rate, item characteristics and scoring method. Practical suggestions have been provided for teachers to design a cloze test. Teachers can develop the reliability and validity of cloze tests by manipulating the above mentioned four factors. However, further research needs to be done to address the following questions.

This study indicates that cultural content has to be taken into account when choosing test material. Background knowledge including cultural knowledge can affect the test result because reading involves top-down processing. Eliminating the cultural content or modifying the original version does not mean that it has to measure only bottom-up processing. A cloze test has to measure top-down processing as well as bottom-up processing. If that is so, what content or which genres has to be selected for cloze material to test both processing needs to be studied more.

Although Kobayashi considered the subjects' first language which does not have articles as a cause for making article items difficult, more research is needed to identify what made the article items difficult because articles are the most difficult to acquire for other language speakers including Japanese according to theories of first and second language development such as Error Analysis and Interlanguage[11].

Reporting that the amount of context is correlated negatively with item difficulty, the above research suggest that the item difficulty should be tailored to the students' level by adjusting the amount of context. However, how to measure the amount of context which is needed to restore an item is still to be answered.

Besides, other alternative scoring methods have to be devised to make the scores by non-native EFL teachers more valid. When there is a spectrum of correct answers according to appropriateness or accuracy, the teachers are apprehensive about making decision on the correct answers, which forces them to choose the exact-word-only method. As a result of that, the scores may lose validity according to the above research. However, applying acceptable-scoring method brings a challenging problem to the teachers though the above research confirmed that

non-native EFL teachers can score a cloze test validly. Therefore delicate ways of helping the teachers who have difficulties in grading or giving partial credit have to be proposed.

In seeking a more valid and reliable cloze format, this research has made some suggestions based on a literature review. We can conclude that a cloze test can be valid only when teachers are aware of the factors which affect the test result and the ways to handle the factors to make the test appropriate for their test purposes.

References

[1] Abraham, R., G., & Chapelle, C., A., The meaning of cloze test scores: an item difficulty perspective. The Modern Language Journal, 76(4), pp.468-479, 1992.

[2] Alderson, J. C., The cloze procedure and proficiency in English as a foreign language. TESOL Quarterly, 13, pp.219-223, 1979.

[3] Bachman, L. F., Performance on cloze tests with fixed ratio and rational deletions. TESOL Quarterly, 19, pp.535-556, 1985.

[4] Bachman, L. F., Language testing-SLA research interfaces. In L. F. Bachman & A. D. Cohen (Eds.), Interfaces between second language acquisition and language testing research, Cambridge: Cambridge University Press, pp.177-195, 1998.

[5] Brown, J. D., What are the characteristics of natural cloze tests? Language Testing, 10, pp.93-116, 1993.

[6] Brown, J. D., Cloze item difficulty. JALT Journal, 11, pp.46-67, 1989.

[7] Chihara, T., Sakurai, T., & Oller, J. W. Jr. Background and culture as factors in EFL reading comprehension. In Oller, J. W. Jr. and Jonz., J. (Eds.), Cloze and coherence, London: Associated University press, pp.135-147, 1989.

[8] Green, B. B., Testing reading comprehension of theoretical discourse with cloze. Journal of Research in Reading 24(1), pp.82-98, 2001.

[9] Han, M., EFL Readers' Test-taking Processes for Completion vs. Multiple Choice Cloze Tests. The Linguistic Association of Korea Journal, 15(3), pp.189-208, 2007.

[10] Huges, A., Testing for language teachers. Cambridge: Cambridge University Press, 2003.

[11] Kobayashi, M., Cloze tests revisited: exploring item characteristics with special attention to scoring methods. The Modern Language Journal, 86(4), pp.571-586, 2003.

[12] Oller, J. W. Jr. Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. The Modern Language Journal, 74, pp.151-157, 1983.

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

[14] Stubbs, J., & Tucker, G .R., The Cloze Test as a Measure of English Proficiency. Modern Language Journal, 58, pp.239-241, 1974.

[15] Yuet, C., Cultural content and reading proficiency: A comparison of Mainland Chinese and Hong Kong learners of English. Language, Culture & curriculum, 16(1), pp.60-69, 2003.

Chung-yeol Park [Regular member]

- Feb. 1992 : Sookmyung Women's Univ., English Literature, MS
- Dec. 2004 : Oklahoma City Univ., TESOL, MS
- Feb. 1999 : Sookmyung Women's Univ., English Literature, PhD.
- Jan. 2007 ~ Dec. 2008 : California State Univ., Dept. of Education, Visiting Professor
- Aug. 2009 ~ current : Korea Nazarene Univ., Owens International College, Professor

<Research Interests>
English Drama, Language Education, English Camp