РАЗДЕЛ IV. ЛИНГВИСТИКА И АКТУАЛЬНЫЕ ПРОБЛЕМЫ ПРЕПОДАВАНИЯ ИНОСТРАННЫХ ЯЗЫКОВ

UDC 811
DOI: 10.18413 / 2313-8912-2015-1-3-104-109

Nancy Burkhalter

PREDICTING STUDENTS' GPAS BASED ON CLOZE TESTS IN KAZAKHSTAN

Nancy Burkhalter
PhD
Lecturer
Seattle University
901 12th Ave, P.O. Box 222000
Seattle, WA 98125
E-mail: nancy.burkhalter@live.com

Abstract

The purpose of this study was to determine what correlation, if any, existed between the MA TESOL entrance exam, specifically the cloze test, and students' performance in the program at an English-medium university in Central Asia. This was accomplished by comparing the results of students' scores on the three parts of the exam (cloze and two essays) with their GPAs. Findings showed the number correct in the second half (with exact-word scoring) of an academic passage had a higher correlation with GPA (0.60; significant at 0.01). Correlations with the essays were not significant. A brief review of the literature of the issues surrounding the cloze procedure relevant to our program are discussed, viz., 1) how well cloze measures ESL reading comprehension, 2) which scoring method correlates with second language proficiency, 3) whether we should score both halves or only the second half of the test, and 4) what the optimum deletion frequency is. The results will be used to further calibrate entrance requirements and exam assessment procedures.

Key words: cloze; reading comprehension; scoring method; deletion ratio.

INTRODUCTION

Our English-medium university in Central Asia began a master's program in Teaching English to Speakers of Other Languages (MA TESOL) in 2007. Given that students in our program, most of whom speak English as a foreign language, must read anywhere from 20 to 50 pages each week from academic texts, accurate evaluation of applicants' reading comprehension is crucial.

To measure reading comprehension, we use the cloze procedure, a test requiring applicants to restore words deleted from a text. In addition to the cloze, other factors are also considered for admittance, such as a university wide entrance exam, two essays on the MA TESOL exam, and an interview to assess spoken English. We take these data and form a profile of the student's abilities, goals, and spoken language proficiency, and depending on their performance, grant regular status, no admittance, or provisional standing. Students falling into the latter category must take remedial English courses (e.g., graduate English classes, critical thinking, and/or foundation English classes) until such time as they are granted regular status.

Since its inception, the program has admitted 41 students and had 16 withdrawals. Of those who withdrew, nine earned a minimum of three credits toward the requisite 44. For purposes of program assessment, we wanted to revisit our testing procedures to determine if they are a reliable and valid method of testing applicants' abilities.

The purpose of this study was to see what correlation, if any, exists between our entrance exam, specifically the cloze test, and students' (N=28) performance in the program. The results of students' scores on the three parts of the MA TESOL exam (cloze and two essays) were compared with their GPAs. The results will be used to recalibrate our entrance requirements and exam assessment procedures.

We were particularly interested in re-examining the issues about cloze and its predictive ability. Cloze has been a controversial method of ESL reading assessment since it began being used by [15]. Proponents of cloze testing assert that it "measures comprehension that ranges beyond the context immediately surrounding a cloze deletion" and see it as "stable, reliable, and sensitive to comprehension processes at various levels." Critics, on the other hand, have found it to be "unreliable and erratic" [8, P. 61]. While there are many issues surrounding cloze testing, the ones that relate to our inquiry include the following: 1) How well does cloze measure ESL reading comprehension? 2) Which scoring method correlates with second language (L2) proficiency?, 3) Should we score both halves or only the second half of the cloze?, and 4) what is the optimum deletion frequency? Following is a discussion of these issues.

OUR QUESTIONS SURROUNDING LOZE

1) How well does cloze measure ESL reading comprehension?

Text comprehension

[12] claims that standardized tests of reading comprehension correlate as high as .80 and above with cloze tests. [1], however, is not sanguine that cloze testing measures overall text comprehension but errs on the side of caution, stating that "...cloze is largely confined to the immediate environment" [P. 225]. After all, he points out, cloze never deletes phrases or clauses, which are, in the main, those elements that carry cohesive devices (with the "exception of anaphora, lexical repetition and logical connectors," [1, P. 225].

While researchers agree that context clues can occur in the near vicinity of the deleted word, and even beyond sentence boundaries, [12] states that some items can be sensitive to constraints as far away as 50 words before or after the blank. Not surprisingly, the more proficient subjects were in English, the more agile they were in reaching beyond sentence boundaries for context clues [4].

Even more difficult for readers are cases where the sense of the meaning is only hinted at through context with no exact referent. Take, for example, the following two sentences. To fill in the blanks using context, the deletions must be understood through a subtle clue requiring close attention to the meaning. (Words deleted are underlined):

Typically, these students come <u>to</u> the first class session <u>feeling</u> both apprehensive and resentful. <u>They</u> are nervous about having <u>to</u> take the class, and, <u>at</u> the same time, they ***<u>suspect</u>*** it will be of <u>no</u> use. [6, P. ix]

If we examine the word *suspect*, it seems at first blush to be a difficult (exact) replacement, but the clue, albeit subtle, is given in the first sentence, stating that students are both "apprehensive" and

"resentful." In the second sentence, the first clause gives an example of their apprehension: because they are nervous about the test taking; the second clause gives an example of their resentfulness: they believe it is not useful. But the word *suspect* is a more appropriate replacement because it is not something the students can confirm (as, say, the words *believe* or *know* might indicate). So in this instance, if readers understand how texts are created, i.e., that statements are backed up with examples, they may be more aware of those clues. But, it may be a nuance that escapes them as they negotiate all the other demands of reading an academic text about a topic they do not know much about in a language and a rhetorical pattern they do not handle with ease, not to mention the anxiety provoked by test-taking in general. (See Appendix for another full passage.)

Context clue aside, [1] cautions that cloze tests can be manipulated both in terms of scoring procedures (discussed below) and difficulty of texts, both of which would alter the expected outcome, and are therefore unreliable.

Alderson is not alone in his skepticism of cloze [9, 10, 14, to name a few. See [8] for a longer discussion of these objections.). But others believe that if testers hew closely to certain rules of test creation and scoring procedures, cloze is indeed reliable and valid and sensitive to comprehension processes at various levels, [e.g., 4, 7, 12].

Also at issue is how reliably cloze measures higher order skills, such as uncovering an author's inferences and assumptions, and other top-down comprehension processes that are largely independent of the words deleted. [8] believes cloze does do that, stating that "the cloze procedure…challenges universal processing mechanisms at all levels from word recognition through concept building; therefore responding to cloze tests must necessarily involve a great deal of higher order language processing" having to do with coherence and cohesion [8].

2) Which scoring method correlates with L2 proficiency?

[2] discusses four possible methods of scoring cloze tests: the exact answer, which accepts only the word from the original text; the acceptable answer, in which any semantically appropriate word for that context is deemed correct; the cloze-entropy, which weights acceptable answers according to their frequency in a pre-test given to native-speakers; and the multiple choice method, which provides a set of alternative answers test takers can choose from to fill in the blank. Brown points out that whereas the first three methods test productive skills, the latter, multiple-choice format, is more likely to be testing receptive skills.

[2] also states that all four methods fare equally well in terms of validity and reliability. If ease of test creation is a high priority, then scoring for the exact word is preferable since it can be scored by anyone. While function words (prepositions, articles, connecting words) require exact-word scoring, [2] finds exact-word scoring for content words "repugnant because it counted an answer wrong, which is actually correct, simply because the author of the original passage did not choose to use that word" (2, P. 316).

We might also add that the cloze procedure assumes the original text was well written. Too, academic prose often uses low-frequency words, thereby increasing the potential for a lower score if the exact-worth method is used. Overall, Brown found the acceptable-word method slightly superior to the exact-word method for testing productive skills when all criteria were considered.

It could also be argued that L2 learners who are able to find an appropriate replacement may be exhibiting greater knowledge of the various definitions and nuances of a word or expression, as well as an overall understanding of the text. This would argue in favor of acceptable-word replacement.

3) Should we score both halves or only the second half of the cloze?

[14] raise the possibility that intersentential comprehension may be greater in the second half of the passages due to accumulated knowledge from earlier text. If this is true, scores on the second half would "prove a better measure of reading, as it allows the measurement of the use of cumulative information, an important skill in the integration of information across sentence boundaries" [14, P. 237]. If it is found that scores on the second half correlate better with our students' GPAs, perhaps we might consider changing policy to reflect that finding, thus foreshortening our grading process.

4) What is the optimum deletion frequency ratio?

A mechanical deletion ratio (e.g., every fifth word) will be insensitive to how many content and function words are removed. Function words

have only one exact replacement and are bound by more immediate context clues than content words, which are recovered from a general understanding of the context, aka intersentential comprehension [14].

[14] cite [13], who gives the rationale for selectively deleting both function and content words: "...function words *might* [emphasis theirs] result in a better measurement of the understanding of structural or syntactic meaning, and that content deletions would result in a better measure of the understanding of the subjective content" [14, P. 236]. Even though deliberately deleting specific words may be a more precise measure of what they call "intersentence integration" [14, P. 237], doing so would add to the burden of the test designers and not worth the effort. Therefore, they conclude, deleting every fifth word is a suitable deletion frequency ratio.

Would placing blanks farther apart aid readers? Not according to [1]. He did find that placing blanks more frequently, say, after every third word, increases difficulty in word restoration because, as [12] states, "much discriminatory power is lost" [12, P. 107]. However, placing blanks after the 6th, 7th, 10th or even 15th word had no effect on predictability of a deletion. So a deletion ratio of every fifth word is preferred.

The length of a cloze test can affect reliability. [3] recommends it be no longer than 50 deleted items. (Our test passages delete about 40 words out of roughly 250 words.) Although longer tests tend to be more reliable than shorter ones, fatigue can set in and students stop taking the test seriously or merely give up. In addition to the deletion rate, other factors that can affect reliability are the students' abilities and the difficulty of the text, as [1] stated above.

RATIONALE

Even though the issues surrounding cloze testing are not settled, it is still very widely used in many areas of education because it is an easy and cost-effective way to administer a reading comprehension test. Since our program is now four years old and about to graduate its first students, we wanted to make our first evaluation of how well the cloze test predicts applicants' performance in our program. (Note: The results of the writing tests are included in this study for purposes of information about our testing procedure but are not the main focus.) Specifically, we wanted to know what correlation(s) exists between the students' MA TESOL test data and their GPAs.

MATERIALS AND METHODS

Participants

GPAs of 28 students who had accumulated at least three credits in our program were compared with their scores on both the written and cloze parts of the MA TESOL entrance test. *Tests*

Writing Samples

Students write for a half hour on each of two essay topics. The first asks them to write about their reasons for wanting to enter the program. The second asks them to respond to a quote stating that learning English in a country where it is spoken is easier than learning it in a country where English is not the first language.

Scoring

Responses are scored using the 6+1 Trait Writing Scoring Continuum developed by [11]. Each essay can earn a maximum of 30 points on the following traits: ideas, organization, voice, word choice, sentence fluency, and convention. (Each trait is worth 5 points; the presentation trait is not scored). The twelve scores from both essays are averaged. A minimum score of 3 is considered passing.

Cloze

The creation and scoring of the cloze test followed the protocol outlined in [14].

Set-up

All the passages from academic textbooks used by our program total about 250 words each. The first and last sentences of the excerpts remain unaffected. Every fifth word is deleted. Students are instructed to read the excerpt and insert the most appropriate word into each blank, with each blank representing only one word. No further instructions are given to the test takers as to how to use context cues or how the passage will be scored.

Scoring

A passing score consists of 50% or better on a minimum of two components of the cloze test. We set the cutoff at 50% because we have been scoring the test using word-accurate insertions. So to compensate, we felt it necessary to be somewhat lenient in the required percentage correct.

The following scores are calculated for each student, using exact replacement criteria:

1. Cloze score: percentage of words replaced for the entire excerpt.

2a. First *half* score: percentage of words correctly replaced in the first half. Exact match is required.

2b. Second *half* score: percentage of words correctly replaced in the second half. Exact match is required.

3a. *Function* words (prepositions, articles, connecting words): percentage correct using exact word replacement.

3b. *Content* words (not function words): percentage correct using exact word replacement.

FINDINGS AND DISCUSSION

DATA ANALYSIS

Each student's cloze scores were summarized by means of descriptive statistics for comparison with the scores of other students. Scores from the writing tests and the cloze procedure were collected from each of the 28 participants.

The cloze test data were recorded as the number correct in each half of the test. The number correct in the second half had a higher correlation with GPA (0.60; significant at 0.01) than the number correct in the first half (0.35, not significant at 0.05). A cloze total score was computed by adding the number correct in the first half and the number correct in the second half. This total score has a correlation of 0.56 with GPA (significant at 0.01).

The essay data consist of six scores for Ideas, Organization, Voice, Word Choice, Sentence Fluency, and Convention. In most cases these scores were averages for two essays (in some cases, only one essay was graded). A total essay score was computed as the sum of the six individual scores. Its correlation with GPA is 0.31 (not significant at 0.05).

The correlation between the total cloze score and the total essay score is 0.39 (significant at 0.05). Although this suggests substantial overlap of the abilities measured by cloze and essay, it should be noted that a small group of students had much better essay scores than their cloze scores would suggest. These students have remained in the program with (mostly) good GPAs.

Little test data were available for those who withdrew from the program. However, it is interesting that of nine who withdrew (and had completed at least 3 credits before doing so), five had GPAs greater than 3.0. It might be worthwhile exploring through interviews why these people withdrew.

Given the small sample size of this study, we can draw only slim conclusions about the predictability of our cloze and writing results and students' success in the program. However, there does seem to be some evidence that 1) students who scored well on the cloze will predict a higher GPA, thus validating our confidence in it as a testing instrument, and 2) the second half of the test correlated with GPA slightly better than the first half, thus lending some support to the notion that the second half measures comprehension any better than the first due to accumulation of knowledge about the passage [14]. However, the data are not robust enough to abandon grading the first half of the test just yet.

This study has led us to change our scoring procedure from exact word only to the acceptable replacement method, as we now feel it is a more accurate indication of our population's reading abilities and will more accurately reflect a student's knowledge of English. This seems to be a prudent decision since an L2 learner's use of a suitable synonym does not in any way mean they do not understand the text. On the contrary, it shows they do understand it and have a word stock large enough to draw from.

Despite the focus on the cloze test in this article, we also found that the reading and cloze may be in effect measuring similar skills; however, there was a small group who attained an average score on the cloze but fared quite well on the essay. This finding also supports our holistic approach to judging applicants' abilities: no one test tells us everything.

Although the data are not robust enough to make major policy decisions about our entrance exam, preliminary information gives us cautious confidence that we are fairly and accurately measuring reading comprehension and predicting students' performance in the program. Moreover, with a few adjustments in our scoring procedure to an acceptable-word method, we hope to show even higher correlations between the cloze score and GPA the next time correlations are run.

CONCLUSION

Using the cloze procedure would also seem a valid measure for educators in any discipline, regardless of the language of instruction, who wish to sample a student's reading comprehension by using a text from their discipline.

REFERENCES:

1. Alderson J.C. *The cloze procedure and proficiency in English as a foreign language*. TESOL Quarterly. 1979. № Pp. 219-223.
2. Brown, J.D. Relative merits of four methods for scoring cloze tests. *MLA Journal*. 1980. № 64. Pp. 311-317.
3. Brown, J.D. *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall, 1996.
4. When are cloze items sensitive to constraints across sentences? Chavez-Oller M., Chihara T., Weaver K., Oller J. Jr. // *Language Learning*. 1985. № 35. Pp. 181-203.
5. Finch G. *Key concepts in language and linguistics*. New York: Palgrave Macmillan, 2005
6. Freeman D., Freeman Y. *Essential linguistics*. Portsmouth, NH: Heinemann, 2004.
7. Jonz J. *Textual cohesion and second language comprehension*. Language Learning. 1987. № 37. Pp. 409-438.
8. Jonz J. *Another turn in the conversation: what does cloze measure?* TESOL Quarterly. 1990. № 24. Pp. 61-83.
9. Kibby M.W. Intersentential processes in reading comprehension. *Journal of Reading Behavior*. 1980. № 12. Pp. 299-312.
10. Klein-Bailey C. A cloze is a cloze is a question. In *Issues in language testing research*, Ed. Oller, J. Jr. Rowley, MA: Newbury. 1983. Pp. 218-228.
11. Northwest Regional Educational Laboratory URL: http://schools.nyc.gov/documents/d75/ais/6plus1traits.pdf (date of access: November 22, 2010)
12. Oller J. Jr., Close tests of second language proficiency and what they measure. *Language Learning*. 1973. № 23. Pp. 105-118.
13. Rankin E. The cloze procedure revisited. In Interaction: Research and practices in college adult reading, Ed. Nacke, P.L. *Twenty-Third yearbook of the National Reading Conference*. 1974.
14. Shanahan T., Kamil M., Tobin A. Cloze measure of intersentential comprehension. *Reading Research Quarterly*. 1982. № 17. Pp. 229-255.
15. Taylor W. "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly*. 1953. № 30. Pp. 415-33.

APPENDIX

Example of a passage chosen for the cloze entrance task where every 5th word is deleted, excluding the first and last sentences.

Instructions: Read the excerpt below. Each blank represents only one word. Replace the deleted words. Write the most appropriate word for each blank.

Phonetics and phonology are concerned with the study of speech and, more particularly, with the dependence of speech on sound. In order to understand the distinction between these two terms it is important to grasp the fact that sound is both a physical and a mental phenomenon... At the same time, however, neither speaking nor hearing are simply mechanical, neutral processes. We endow the sounds we make and hear with meaning. They have a mental or cognitive existence as well as a physical one. Another way of putting this is to say that sounds are psychologically, as well as physically, real. Psychological reality is important in linguistics. Sometimes things can be psychologically real without having any real-world correlates. So, for example, most people will idealize their own speech and hear themselves speaking perfectly clearly and accent-neutral, when in fact the reverse is the case.

This division between the physical and mental dimensions of speech sounds is reflected in the terms 'phonetics' and 'phonology'. Precisely where the division comes has been, and still is, a matter of fierce debate between phoneticians and phonologists... Phonetics is really a technically based subject concerned with measuring sound, recording frequencies, and generally studying the physiology of speech. Phonology, on the other hand, is essentially preoccupied with sound as a system for carrying meaning. Its fundamental concern is with identifying phonemes. These are the small building blocks of the spoken language that provide the skeleton framework of speech. [5, P. 32]