Interpreting Cloze Scores in the Assessment of Text Readability and Reading Comprehension

Graham Wagner

In the face of the confusion in recent literature on the subject of interpreting cloze scores it appears necessary to draw attention to the seminal research of John Bormuth, especially his 1971 study on cloze criterion passage performance. It was in this study that Bormuth convincingly and empirically demonstrated the need to replace the traditional multiple-choice tests of reading performance as the basis of comparison for the cloze, and substitute a more valid frame of reference by which scores on cloze tests might be more meaningfully interpreted. Even so we still find some authors (Harrison, 1980, for example) adding to the confusion of appropriate criterion level cut offs, rather than resolving it.

In the early days of cloze reading comprehension and readability research, researchers attempted to set criterion standards of performance based upon multiple choice comprehension test scores primarily in relation to the well-known, and generally accepted, 75% and 90% levels of reading comprehension. With the discovery of the cloze procedure by Taylor (1953) it has been said that "performance criteria should be established that would allow teachers to use the simple and objective cloze procedure, rather than the objective, problem-ridden multiple choice tests" (Boyce, 1974).

Bormuth (1967) was the first researcher to establish cloze passage performance criteria in relation to multiple-choice comprehension test scores. In his study he compared the performance of 100 pupils in Grades 4 and 5 and concluded that if the widely agreed upon comprehension levels criteria were accepted then a score of 38% on the cloze test was equivalent to 75% on a multiple-choice test, while a 50% score on a cloze was equivalent to 90% of the multiple-choice test (over the same material). When corrected for guessing the equivalent scores were 43% and 52%.

In the next year Bormuth (1968) undertook another study to determine relative criterion performance scores, this time using graded paragraphs from the Gray Oral Reading Tests (1963). Employing a simple matching

procedure to determine comparable scores Bormuth found that the resulting cloze equivalents of the criterion performance scores of 75% and 90% respectively were 44% and 57%. Bormuth at the time argued that the difference between his 1967 and 1968 studies were due to the former incorporating a ceiling effect on the multiple-choice scores at the upper end of the range, resulting in lower cloze scores. Boyce (1974) adds that these differences may be partly due to the methods Bormuth used to obtain equivalence.

Following on from the early work of Bormuth, Rankin and Culhane (1969) sought to replicate Bormuth's 1967 study. Using only fifth grade children as subjects these researchers found that cloze scores of 41% and 61% respectively were equivalent to multiple-choice reading test score of 75% and 95%.

Because of the lack of agreement among researchers about meaningful and empirically sound criterion levels for interpreting cloze scores, Bormuth (1971) undertook a study of rational cloze criterion passage performance in the US Department of Education *Final Report,* which, it appears, is seldom read by researchers. This paper seeks to focus attention on Bormuth's 1971 work, not only to clear up a misunderstanding about the dubious validity of using multiple-choice tests as criterion measures in cloze interpretation but to standardise interpretation as far as can be done by taking into account systematic and random error in educational achievement measurement.

Reading as information gained

While it is particularly difficult to define reading with understanding in terms of a process (because we do not know what is actually going on in the minds of those who are reading) it is possible to talk about comprehension as a "set of generalised knowledge acquisition skills which permit people to acquire and exhibit information gained as a consequence of reading printed language" (Bormuth, 1971).

This definition of reading has, as Hansen and Hesse (1974) point out, several advantages when constructing a model for assessing reading comprehension: "First, information gained can be defined behaviourally, second, information gained can be defined in the context of the specific kinds of materials... third, the definition skirts an important dilemma

i.e. the inability at the moment to adequately define the reading process" (pp. 7-8).

The same researchers go on to summarise the arguments used against multiple-choice tests by Bormuth. To measure reading as information gained, they say, we might employ traditional multiple-choice tests such as are used in most standardised reading comprehension tests. These tests, however, have serious limitations for:

> they allow considerable guessing (at least 20-25%) and cueing; they can be made easier or harder by judicious selection of the incorrect options (distractors) by the test item writer; and they often introduce new words and phrases not found in the original written message. A student may get a test item wrong not because he did not understand the test passage, but because he did not understand some new word or phrase in the test item. Such test items are said to lack objectivity. Moreover, such traditional approaches must be rejected for they simply measure how many questions can be answered after the passage has been read. There is no indication of how many questions could have been answered prior to having read the selection. In other words, there is no measure of information gained from the printed passage (Hansen and Hesse, 1974, p. 8).

Bormuth (1971) used a model for measuring "information gained" which avoids the problems associated with using multiple-choice tests of reading comprehension. A reading of his report will give a detailed explanation of how his model works. Suffice it to say that his information gain tests had several ideal qualities:

> First, the test items, because they are a random sample of all possible test items for a given passage, reflect the actual relative difficulty of the passage within the bounds of sampling error. The test items have *not* been tampered with by the item writer to make them easier or harder than the content material should actually reflect. Second, because they require the subject to recall and respond they reduce guessing to a minimum. Finally, such items can be administered before and after reading the passage without the confounding effects of guessing (Hansen and Hesse, 1974, p. 14)

Thus, it is concluded by Hansen and Hesse, the information gained model is an excellent way of dealing with the assessment of reading proficiency except that these tests are expensive and time-consuming to make, and require the services of a linguistics expert to produce the "generative question" test items.

The cloze

Bormuth in his aforementioned research discovered that the cloze was systematically and closely related to the measure of "information gain". Using regression procedures (trend analysis), Bormuth demonstrated that student performance on tests of "information gain" tracked directly with student performance on the cloze versions of the same tests. Furthermore, using the same basic design in five major studies, and at grade levels 4 through to 12 in different schools, he was able to show that a score of 35% "would appear to be a defensible criterion of the threshold of literacy when applied to cloze scores of specified subjects on specific materials" (Hansen and Hesse, 1974, p. 17). In other words, once students get to 35% on a cloze test they have moved out of the 'frustrational' level into the 'instructional' level of reading comprehension.

In further complementary studies Bormuth (1971) included other important reading related variables such as (a) reading age, (b) interest in the materials, (c) purpose of the reading material and (d) willingness to read the material, to arrive at a criterion score for the threshold of 'independent' reading. Taking into account all factors involved in Bormuth's research it could be said that "... 50% might be posited as a criterion of comfortability, that point above which the student will be able to gain information from the material with some ease and comfort" (Hansen and Hesse, 1974, p. 19).

The result of Bormuth's 1971 research means that in terms of assessing reading comprehension and reading materials that are used in primary and secondary schools, a cloze score range of 0 to 34% is equivalent to the 'frustrational' level; 35 to 49% is equivalent to the 'instructional' level and 50% and above is equivalent to the 'independent' level. Bormuth, also indicated, in the aforementioned report, that within the range of students that he studied, the criterion level varied somewhat according to whether the material was reference material, textbook material or recreational reading material. Therefore, if a teacher was to check the readability of a class textbook he might expect the cloze test to show that his students are spread between the 'instructional' and 'independent' levels. On the other hand, a suitable recreational book would probably show that his students were mainly in, or bordering on, the independent reading level although in most classes, because of the wide range of abilities, one would also find some students at the frustrational level as well.

In summary then, Bormuth's 1971 extensive research to establish the cloze criterion passage equivalents to the frustrational, instructional and independent reading levels has now given teachers and researchers a common standard by which to judge student reading performance. Most of the confusion among researchers and writers about the appropriate criterion levels is due to the fact that most people have not read Bormuth's paper (or later works) probably because of the complexity of the technical arguments and mathematical procedures used. Even so, Bormuth succeeded in showing that there is a better, more valid and reliable way of interpreting cloze score performance than by equating student reading performance to equivalent scores based upon multiple-choice test criteria. While it is not suggested in this paper that Bormuth's criterion levels are the final word, they are theoretically and empirically the soundest basis upon which to interpret cloze test performance. That is, of course, as long as the cloze tests have been developed according to the procedure laid down in Elley (1977) and Gilmore and Wagner (1985). For as in all tests, if the test construction is faulty, then the results will also be faulty.

References

Bormuth, John R. (1967) *Implications and Use of the Cloze Procedure in the Evaluation of Instructional Programmes.* Occasional Report No. 3, Centre for the Study of Evaluation of Instructional Programmes, Los Angeles: University of California.

Bormuth, John R. (1968) 'The Cloze Reliability Procedure.' *Elementary English* 45, 429-436.

Bormuth, John R. (1971) 'Development of Standards of Reliability: Toward a Rational Criterion Passage Performance.' *Final Report,* Project No. 9-87. Year of Research, US Office of Education.

Boyce, (1974) Some Difficulties in Using Cloze Procedures to Assist Readability. Master of Education Thesis, University of Melbourne.

Elley, Warwick (1977) 'Hanging Out the Cloze.' Set 4. Wellington: NZCER.

Gilmore, A. and Wagner, G. (1985) *The Readability of Trade Examinations.* Wellington: NZCER.

Grey, W.S. (1963) *Reading Tests.* Indiannapolis: Bobbs-Merrill.

Hansen, Lee H. and Hesse, Carl D. (1974) *A Pilot Literacy Assessment of Madison Public School Students.* Final Report, Department of Research and Development, Madison, Wisconsin Public School.

Rankin, E. and Culhane, J. (1969) 'Comparable Cloze and Multiple Choice Comprehension Test Scores.' *Journal of Reading,* 13, 193-198.

Taylor, W. (1953) 'Cloze Procedure: A New Tool for Measuring Readability.' *Journalism Quarterly,* 30, 415-433.