THE IMPLICATIONS AND USE OF CLOZE PROCEDURE IN THE EVALUATION OF INSTRUCTIONAL PROGRAMS

John R. Bormuth

CSEIP Occasional Report No. 3, April 1967
University of California, Los Angeles

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

The Implications and Use of Cloze Procedure in the Evaluation of Instructional Programs

John R. Bormuth

One purpose of this paper is to examine the utility of the cloze readability procedure as a device for evaluating the effectiveness of instructional programs. The cloze readability procedure consists of a set of rules for selecting samples of verbal text from written instructional materials and for making, administering, scoring, and interpreting cloze tests made from those samples. In its essential form, the cloze readability procedure purports to be no more than a method for determining the extent to which students understand the instruction they receive from the written verbal material.

Methods of the type represented by the cloze procedure are presently essential to the process of evaluating instructional programs. Evaluations which include only a measure of the outcomes of a program and a judgment of their worth ignore the fact that the knowledge taught in an instructional program is selected in competition with other knowledge that is also valued. One of the most painful realitites of a curriculum construction is the fact that much valued knowledge must be excluded because there are not sufficient time, money, and other resources to permit its inclusion. For this reason, the evaluation of an instructional program cannot be considered complete unless it assesses the efficiency of each of the components of the instructional program. The cloze readability procedure has been developed to provide information

of this kind.

This is not to say that the cloze readability procedure or any procedure in which materials are tested directly on the students represents the ideal approach to assessing the efficiency of instructional materials. Such procedures give the evaluator an indication of how much learning results from the exposure of students to the materials, but give him little information about how the various features of the materials influenced that learning. If a mature science of instruction existed, it would be possible, merely from an examination of the features of the instructional materials, to calculate the kind and amount of influence any given feature or set of features would exert on the outcomes. Indeed, this might be said to be the ultimate objective of much of the research in school learning. Until this objective has been achieved, expedients such as the cloze readability procedure must play a vital role in the evaluation of instructional programs that utilize written, verbal instructional materials.

The second purpose of this paper is to examine the possibility of developing a method, which incorporates the cloze procedure, for making criterion reference tests over verbally presented instruction. It is not possible to evaluate the outcomes of an instructional program unless there is some way to determine what content was taught and whether each item of content was learned. Conventional test making procedures offer no method for objectively deriving a list of the items of content taught. If such a list could be derived, conven-

tional test writing theory offers no objective procedure for deriving test questions from the list of content items. Because there is no rigorous way to determine whether the knowledge measured by a test is representative of the knowledge taught by the program, there is clearly no way to construct a criterion referenced test over verbally presented instruction.

In the second section of this paper, test item writing theory will be cast into a more definite form than it has taken in the past. A procedure will then be proposed for making criterion reference tests from programs containing verbally presented instruction. It will be seen in the course of this discussion that the use of the cloze procedure is essential for the selection of the items included in criterion referenced tests made from verbally presented instruction.

Cloze Readability Procedure

Cloze tests can be made in a variety of ways. When they are used to measure the comprehension difficulties of text materials, however, investigators almost invariably use a specific set of procedures called the cloze readability procedure. Cloze readability tests are made by deleting every fifth word from a passage. The deleted words are replaced by underlined blank spaces of a uniform length, and the tests are mimeographed.

Cloze readability tests are given to subjects who have not been permitted to read the passage. The subjects are

instructed to write in each blank the word they think was deleted to form that blank. A response is scored "correct" when it exactly matches the word deleted. The difficulty of a passage is the mean of the subjects' percentage scores on the test.

The difficulty of every word, phrase, clause, or sentence in the passage can also be determined by using five forms of a cloze test over the passage. To make the first form, words 1, 6, 11, etc. are deleted; words 2, 7, 12, etc. are deleted to make the second form. This procedure continues until all five forms have been made and every word in the passage appears as a cloze item in exactly one test form. The proportion of subjects writing the correct word in a blank is used as a measure of the difficulty of the word deleted. The difficulties of the words within a phrase, sentence, or passage are averaged to determine the difficulties of those units.

Other Evaluation Methods

Readability Formulas. Perhaps one of the chief reasons why instructional materials are not routinely evaluated to determine whether they have a suitable level of difficulty is that there has been no technique that is at once convenient, economical, and valid. Readability formulas are convenient, inexpensive, and require only unskilled clerical assistance to use, but the formulas currently available have validities that range from .5 to only about .7. Further, the equations

take into account only a limited range of linguistic variables and the variables that are taken into account are, by today's standards, crude. Recent research by Coleman (1966a) and Bormuth (1966a) shows that readability formulas having high validities can be developed, but the research that will obtain these formulas is still in progress.

Direct Testing. Using conventional comprehension tests to test materials directly on students seems more valid than using readability formulas, but it is also expensive and unreliable. Because the test items themselves represent a reading task for the student, it is uncertain whether it is the difficulty of the passage or the difficulty of the items that is measured by this procedure.

Programming. Instructional programming might be said to be a third method of determining the difficulty of materials. As programming is currently done, it is an expensive process. Further, programming techniques employ test items similar to those used in conventional comprehension tests. As a result, the criticisms leveled at the use of conventional comprehension tests hold also for programming.

Validity of Cloze Readability Tests

If cloze readability tests are to be used as a measure of the comprehension difficulty of written instructional materials, evidence showing that the tests measure the reading comprehension abilities of students is needed. Further, it

must be shown that the difficulties of cloze tests correspond to the difficulties of other tests used to measure the difficulty subjects have in understanding materials.

Criteria of Validity

Two Concepts of Comprehension. It is necessary to analyze the concept of comprehension further, since there is a fundamental disagreement about which of two measurement operations best represents the concept of comprehension ability. Traditionally, the comprehension ability of a person is measured by having him read a passage and then testing his knowledge of the content of the passage. But scores derived in this manner measure both the person's knowledge acquired as a result of reading the passage and the knowledge he possessed before he read the passage. Comprehension measured in this way will be referred to as post-reading knowledge. On the other hand, many hold that comprehension ability is a set of generalized skills enabling the individual to acquire knowledge from materials. Reasoning from this point of view leads to the claim that comprehension ability is best represented by a score obtained by finding the difference between scores on a test administered before and after the passage is read. Comprehension measured in this way will be referred to as knowledge gain.

Value Placed on Both Concepts. Both conceptualizations of comprehension are relevant to the evaluation of instruction-

al materials. Of course it is highly desirable to select materials from which students acquire much new knowledge. Despite this, previously acquired knowledge is deliberately included in materials in order to provide the repetition essential for retention and in order to state the relationships between knowledge previously acquired and the knowledge being presented for the first time. Hence, a measure used to assess the comprehension difficulty of materials should, ideally, be capable of measuring comprehension in either or both of these ways, since both represent desirable characteristics of materials.

Validity Research

Measurement of Post-Reading Knowledge. Nearly all the validity research on cloze readability tests has concentrated on demonstrating their validities as measures of post-reading knowledge. It seems that only one study approached this problem experimentally. Bormuth (1962) made a cloze and multiple choice test over each of nine passages. The passages were written so that they varied systematically in subject matter and language complexity. Both sets of tests were given to subjects in grades 4, 5, and 6. Each of the main effects and the interaction between language complexity and subject matter produced significant and roughly proportionate effects on the cloze readability and multiple choice scores.

A large number of studies have reported correlations between cloze readability test scores and scores on tests of the type to which the label "comprehension" is conventionally applied. The first studies discussed used comprehension tests made from the same passages as the cloze tests. Taylor (1956), using Air Force trainees as subjects, found a correlation of .76; Jenkinson (1957), using high school students, found a correlation of .82; Bormuth (1962), using elementary school pupils, found correlations ranging from .73 to .84; and Friedman (1964), who used college students, gave comprehension tests consisting of 8 to 12 items each and obtained correlations ranging from .24 to .43. These correlations seem high in view of the fact that, where test reliabilities were reported, the validity correlations and the reliabilities were approximately of the same magnitude.

A fairly large number of studies have reported correlations between cloze readability tests and standardized tests of reading achievement. Table 1 shows the studies and the correlations reported. It is difficult to interpret these correlations because the authors frequently failed to report the variances and reliabilities of the tests for the subjects used in their studies. This was a problem especially in the studies using college students. College students could be expected to exhibit a curtailed

Table 1

Correlations Between Cloze Readability Tests and Standardized Tests of Reading Achievement

Study	Subjects	Tests	Correlations
Jenkinson (1957)	High School	Cooperative Reading C2	
		Vocabulary	.78
		Level of Comprehension	.73
Rankin (1957)	College	Diagnostic Survey	
		Story Comprehension	.29
		Vocabulary	.68
		Paragraph	.60
Fletcher (1959)	College	Cooperative Reading C2	
		Vocabulary	.63
		Level of Comprehension	.55
		Speed of Comprehension	.57
		Dvorak-Van Wagenen	
		Rate of Comprehension	.59
Hafner (1963)	College	Michigan Vocabulary Profile	.56
Ruddell (1963) (6 cloze tests)	Elementary	Stanford Achievement	
		Paragraph Meaning	.61-.74
Weaver and Kingston (1963, 2 cloze tests)	College	Davis Reading	.25-.51
Green (1964)	College	Diagnostic Reading Survey	
		Total Comprehension	.51
Friedman (1964) (20) cloze tests)	College (Foreign Students)	Metropolitan Achievement	
		Vocabulary	.63-.85
		Total Reading	.71-.87

distribution of individual differences which would reduce the sizes of the correlations. But, when this fact is taken into account, the correlations shown in Table 1 seem reasonably high.

Two studies investigated the factor validities of cloze tests. Weaver and Kingston (1963) performed a principle component analysis on the correlations among various tests. The tests included some classifiable as cloze readability tests; they also included a standardized test of reading comprehension. The cloze tests exhibited low correlations with the principal component with which the comprehension test had its highest correlation. Bormuth (1966b) pointed out that this study contradicted the findings of much of the earlier research on cloze tests. In brief, he showed that the correlations involving other tests in the battery exhibited correlation patterns that were highly unusual for them, and that the population of subjects exhibited a curtailed range of variability. He then presented an analysis of data from an earlier study (1962) which showed that a single component accounted for nearly all the variance in a set of cloze tests and multiple choice comprehension tests.

Measurement of Knowledge Gain. There is still only a small amount of information bearing on the question of whether cloze tests are useful as measures of knowledge gain, and this scant information is indirect. Taylor (1956) and Rankin (1957)

each found that subjects who read the intact passages before taking the cloze tests made from those passages achieved higher scores than subjects who had not read the passages. On the other hand, Green (1964) found that having subjects read the passages before taking the cloze tests did not increase their cloze scores over the scores they achieved on a cloze test given them before they read the passage. Rankin (1965) challenged Green's results pointing out that Green failed to correct for the regression effects present in studies using this design.

Measurement of Passage Difficulty. A reasonably substantial amount of research has accumulated showing that cloze readability test difficulties correspond closely to the difficulties of passages as measured by other methods. Taylor (1953), the originator of the cloze procedure, found that cloze readability test difficulties ranked the passages in the same order in which the readability formulas ranked them. When he selected three additional passages which, when judged subjectively, ranked one way but, when analyzed by readability formulas, ranked in the reverse order, the cloze readability test difficulty rankings agreed with the subjective judgments. Sukeyori (1957) found a correlation of .83 between the combined subjective rankings given eight passages by three judges and cloze readability test difficulties of the passages. Bormuth (1962) found a correlation of .92 between the cloze

readabilities of 9 passages and the difficulties of multiple choice comprehension tests made from the same passages. In a more recent study Bormuth (1966) used four sets of 13 passages each and found correlations ranging from .91 to .96 between the cloze readabilities and the comprehension difficulties of the passages. The correlations between the mean number of words pronounced correctly by subjects who read the passages orally and the cloze readabilities of the passages ranged from .90 to .95.

Cloze Test Reliability. When cloze readability tests are used only as measures of the relative abilities of subjects, they are probably somewhat less reliable than well-made multiple choice tests containing the same number of items. For example, Bormuth (1962) found that the reliabilities of the nine, 31 item, multiple choice tests used in his study exhibited reliabilities about equal to those of the nine, 50 item, cloze readability tests made from the same passages. It seems likely that this may have resulted from the fact (Fletcher 1959 and Bormuth 1962) that cloze readability tests nearly always contain a number of very difficult and very easy items which are less efficient discriminators (Davis 1949) than items in the intermediate range of difficulty. However, the large number of very difficult and very easy items appearing in cloze readability tests is actually an asset, making the tests useful in testing subjects differing widely in ability. Zero scores, maximum scores,

and skewed distributions are rarely observed when cloze readability tests are carefully administered. But this range apparently has its limits. Gallant (1964) found that cloze test reliability was reduced sharply when the tests were used with first grade children.

Application of the Cloze Readability Procedure

A substantial body of research has dealt with the technical questions arising when cloze readability procedure is used to evaluate the difficulty of instructional materials. The results of this research seem to justify the application of the procedure to a range of evaluation tasks. The following discussion considers the major problems encountered at each step and discusses the research dealing with those problems.

Designing the Testing Procedure

Cloze readability procedure may be adapted either to measuring the difficulties of short or long passages or to measuring the difficulty of a given piece of material for an individual or for a whole group. Because the number of possible testing designs are almost infinite, only three designs will be discussed to illustrate the principles and problems of designing materials evaluation studies.

Multiple Sampling Problems. When the cloze readability procedure is used to determine the difficulty of a text, the

investigator often deals simultaneously with three samples. First, because it is often impractical to test materials on the whole population with whom the materials are to be used, the investigator draws a sample of pupils to represent this population. The accuracy of his results depends, in part, on the extent to which the sample is representative of the population.

Second, the items in a cloze test represent only a sample of the items that can be made over that passage. When long texts are evaluated, it may be an inefficient use of resources to make all five of the cloze test forms over the passages studied. Therefore, the investigator must sometimes deal with what is called item sampling error. The Kuder-Richardson (1937) formula 21 for calculating test reliability takes item sampling error into account (Lord 1955). The error of the mean that is due to item sampling error may be usefully estimated by Lord's (1955) formula 21. A less complicated procedure is to use two or more cloze test forms over the same passage, and then calculate the variance of the form means. Subtracting the population sampling error variance from the variance of the form means gives an estimate of the item sampling error.

Third, when a lengthy text is evaluated, it is generally not practical to make a cloze test over its full extent. In consequence, sample passages must be drawn from the text and

the cloze tests made over just the sample passages. Hence, the investigator must consider passage sampling error. Passage sampling error can be estimated by finding the difficulty of each of the passages in the sample, calculating the variance of the passage difficulties, and then subtracting the population and item sampling error variances.

Designs. An elaborate design for a text evaluation study might follow these steps. First, the sections of the text are numbered consecutively and passages drawn randomly from each chapter. Two or more passages are drawn from each chapter so that the relative difficulties of different chapters can be compared. Second, two or more forms of a cloze test are made from each passage. The tests should be nearly identical in the number of items they contain. Third, the sample of pupils is drawn randomly, or as nearly so as possible, from the population with whom the cloze tests are to be used, and each pupil is randomly assigned to take one of the cloze tests. When two or more texts are being evaluated, this design permits the investigator to use analysis of variance to ascertain whether the materials differ significantly and to determine how variable each text is from chapter to chapter.

A less expensive procedure consists of using shorter passages---passages of about 50 words. Two forms of a cloze test are made from each passage and the passages are formed

into a single test having two forms. The tests are then given to pupils drawn randomly from the population. This procedure also permits the comparison of two or more different texts. It does not, however, permit the comparison of chapters within a text. It is also less reliable because shorter passages were used.

The simplest problems are presented by the evaluation of short passages such as test items, picture captions, and other passages of less than 1,000 words. All five forms of a cloze test are made from the passage and each form is given to a different, randomly-selected sample of pupils. Where the passage is very short, (containing fewer than 30 items), it is doubtful that individual scores are sufficiently reliable to permit an accurate judgment of how well a given individual understood the passage. The results do provide an accurate estimate of how well the group as a whole understood the passage.

Problems. The first problem encountered is deciding how many pupils, cloze test items, and sample passages should be used. Increasing the number of each reduces the error in estimating the difficulty of the materials, but by different amounts. Bormuth (1965a) found that increasing the number of items in a cloze test reduces error more rapidly than adding the same number of students. There is, at present, no data on the relative size of the error resulting from pas-

sage sampling. The second problem stems from the conjecture that the difficulty of a sample passage from a text may depend, in some degree, on whether the pupil has studied the text preceeding the passage. While this may present little problem in most content areas, it is conceivable that in areas such as the science, the effect could be considerable. This would seem to indicate that some evaluation studies should be designed to accompany instruction in such a way that the pupil is tested on a passage just before he is to study the section containing that passage.

Deletion Procedure

While nearly all readability research employs tests made by deleting every fifth word, cloze tests can be made by deleting every nth word, words at random, or just the words of a given type. The only restriction is that the words deleted must be selected entirely by an objectively specifiable process, otherwise the test must be classified as a common completion test (Taylor 1953).

Cloze test users encountered the problem of discovering how many words of text had to be left between cloze items. Leaving fewer words between items makes it possible to obtain a larger number of items from a given length of text and reduces the number of test forms that

have to be made in order to eliminate item sampling error. Leaving too few words between items, by contrast, introduces the possibility that items will exhibit statistical dependence of the sort where the probability of a subject responding correctly to an item is dependent upon whether or not he is able to answer adjacent items. When appreciable statistical dependence exists, test scores cannot be treated by conventional statistical methods. MacGinitie (1961) studied the problem by varying the number of words of text left intact on either side of a set of cloze items. He was unable to detect any dependence among items when four or more words of text were left between items.

Taylor (1953) pointed out that methods involving the deletion of only words belonging to certain categories had to be excluded for use in readability studies because the frequency with which such words occur in a passage may itself be a variable influencing the difficulty of the passage. There seems to have been no research dealing with some of the more technical problems in the deletion process such as the problem of what should be deleted when a numeral is encountered. For example, should 128 be treated as if it contained three words or should it be deleted as a unit? It is not even clear if a criterion can be found for deciding issues of this sort.

Test Administration

The two principle alternatives in administering a cloze test are to give it either to subjects who have not read the passage or to subjects who have first been exposed to the passage. Giving the cloze test to subjects who have not read the passage obviously uses time more economically. Moreover, it might be argued that giving a cloze test to subjects after they have read the passage causes scores to be influenced by the subject's rote memorization of the passage. (Rote memory is a process commonly held different from comprehension).

The results of validity studies indicate that it makes little difference which method is used. For example, Taylor (1956) found that scores on cloze tests administered after subjects had read the passages exhibited both slightly greater variances and slightly higher correlations with comprehension tests than cloze tests administered to subjects who had not read the passage. Rankin's (1957) studies showed the same results. The greater variance alone seems sufficient to account for the increased correlation. Consequently, when greater validity or reliability is desired, it is probably more economical to obtain it by increasing the number of items in the cloze test and by giving the tests to subjects who have not read the passage.

Scoring Procedure

A response can differ from the deleted word in semantic

meaning, grammatical inflection, and spelling. Users of cloze readability tests nearly always score "correct" just those responses where the stem of the response, the uninflected form of the word, exactly matches the word deleted. The research seems to support this practice. Taylor (1953) found that scores obtained by counting synonyms in addition to responses exactly matching deleted words were no better than scores obtained by counting only responses exactly matching the words deleted when the scores were used to discriminate among passage difficulties. Rankin (1957) and Ruddell (1963) found that scores obtained by counting words exactly matching and synonyms of the deleted words resulted in the scores having slightly, but not significantly, greater variances and correlations with scores on comprehension tests.

In the past, some investigators scored responses "correct" when they were inflected differently from the deleted word. Bormuth (1965b) studied the correlations between comprehension test scores and several categories of cloze test scores which were obtained by counting responses classified according to whether their inflections were correct in the context of the blank and further classified according to whether the stem of the response exactly matched, was synonomous with, or semantically unrelated to the deleted words. All scores obtained by counting grammatically correct responses exhibited positive correlations. The correlation

involving a count of exactly matching responses was .84; the one involving a count of synonyms was .64; and the one involving semantically unrelated responses was .56. All other correlations were either negative or so small as to be indistinguishable from zero. Further, a multiple regression analysis showed that scores based on a count of the responses which exactly matched the deleted words in both inflection and word stem accounted for 95 per cent of the comprehension test variance that could be predicted from the total set of cloze test scores. Thus, it would seem that the most economical and objective method of scoring cloze tests, the exact word method, yields the most valid results.

Most investigators score misspellings correct when the response is otherwise correct and when the misspelling does not result in the correct spelling of another word that also fits the syntactic context of the blank. No research seems to have tested the validity of this practice. Similarly, the influence of illegibly written responses has not received study.

Interpretation of Scores

The difficulty of a text should be reported in terms that make clear how appropriate the text is for a given individual or group. This may be accomplished either by stating

the proportion of the group which is able to achieve cloze readability scores at or above some criterion level of performance or by stating the level of achievement possessed by pupils who are able to attain the criterion level of performance. To do either requires that a criterion score on cloze readability tests be established as representing an acceptable level of understanding a passage.

Criterion Score. Establishing a criterion of acceptable performance on a cloze readability test presents two major problems. First, since cloze readability tests have been in use for only a short time and since they differ radically in difficulty from conventional tests, users have not yet developed a "feel" for what is acceptable performance on a cloze test. Second, the establishment of a criterion score has traditionally been viewed as a matter to be left to personal preference or arbitrary choice rather than as a matter for rational decision based, at least in part, on empirical data.

The most direct approach to establishing a criterion score for cloze readability tests is to adopt a criterion score traditionally used and then to determine what cloze score is comparable to this criterion score. Bormuth (1966c and 1966d) adopted the 75 per cent criterion score which has a long tradition of acceptance (Thorndike 1917) and wide spread use in current practice (Betts 1946 and Harris 1962). According to this criterion, a passage is said to be suit-

able for use in a pupil's instruction if he responds correctly to 75 per cent or more of the questions asked him about the passage. In one study, Bormuth used multiple choice tests and had the pupils read the passages silently. In the other study using different materials and subjects, he used short answer completion tests and had the pupils read the passages and respond to the questions orally. In both studies a cloze score of about 44 per cent was found to be comparable to the 75 per cent criterion. Since the exact word method of scoring was used in both studies, this cloze criterion score is useful only for interpreting other cloze readability tests scored according to that method.

A more adequate approach to the establishment of a criterion score was demonstrated by Coleman (1966b) who set out to determine what level of passage difficulty resulted in the greatest amount of information gain on the part of students reading the passages. He measured information gain by typing the passage on a transparency and covering the words with strips of tape. When this was projected, the student was asked to guess and write down the first word. That word was then exposed and the student was asked to guess the next. Following the first run through the passage, the tape was replaced and the procedure repeated. The difference between a student's scores on the two trials was taken as a measure of information gain. Passage difficulty was determined on a matched group of subjects using cloze read-

ability tests. Interestingly enough, his results seemed to show that maximum information gain occurred on passages having difficulties of close to 44 per cent, the cloze score found to be comparable to the traditional 75 per cent criterion. A question has been raised (MacGinitie 1966) about whether the "information gained" by the subjects in Coleman's study was influenced unduly by rote memorization. Whatever the merits of that conjecture, it seems clear that Coleman's study demonstrated how a rational approach can be made to the establishment of criterion scores.

Reporting Passage Difficulty. The simplest method of reporting difficulty scores is to report the mean difficulty of the text and the proportion of subjects whose score exceeded the criterion score. However, this method limits the general usefulness of the results. It is often impossible to draw the subjects in such a way that they are a representative sample of the pupils with whom the materials are to be used. There is no way to be sure, therefore, that the proportion of subjects who reached the criterion score in the sample will represent the proportion in the population. And, even if the sample of subjects were representative of the population in a school system, it is virtually certain that the sample would not be representative of the subjects in the total population of pupils with whom the materials are to be used. Since text readability studies are of general interest and since they are somewhat costly to conduct,

it seems advisable to use a somewhat more generally useful method of reporting the difficulty of a text.

A fairly easy method to use results in giving a grade placement number to the text. First, the subjects' scores on the cloze readability tests are correlated with their scores on a test of reading achievement; then, using the regression prediction formula, the achievement grade placement score that corresponds to the cloze readability criterion score is calculated. The grade placement score can then be interpreted as the average achievement of subjects who were able to attain the criterion level of performance on the cloze tests made from the text. Other schools using the same achievement test can estimate the appropriateness of the text for their pupils by determining what proportion of the pupils have achievement scores that exceed the reported passage grade placement. Further, since there are many published studies of the comparability of achievement test norms, the results should be useful almost regardless of what achievement test a school uses.

Conclusions

The use of the cloze readability procedure seems to result in valid measurements of the comprehension difficulty of written instructional material. The correlations between cloze readability and conventional comprehension test scores

are high, and none of the research has presented convincing evidence that the processes employed in responding to cloze readability tests are, in any major sense, distinguishable from those employed in responding to conventional comprehension tests. Moreover, passage difficulties determined using cloze readability tests correspond closely to the passage difficulties obtained using other measures.

The cloze readability procedure has a number of advantages not shared by other available methods of determining difficulty. Unlike the conventional test items used in other methods where materials are tried out directly on students, cloze test items are easily made and do not inject irrelevant sources of variance into the measurement of difficulty. Further, cloze readability procedure yields far more valid results than the readability formulas presently available. However, when the readability formulas, now in developmental stages, become available for general use, they will probably be almost as valid and much less costly to use than the cloze readability procedure.

Research on the technology of the cloze readability procedure seems sufficient to permit the application of this procedure to a wide range of materials evaluation tasks, but three important problems remain to be solved. First, it is not at all certain if cloze readability tests can be used to measure knowledge gain. Second, a criterion level

of performance has yet to be established on a rational basis. Third, it still must be determined if the act of isolating a passage from its context affects the difficulty of the passage. A few other problems are also unsolved. For instance, there is the question of how to handle numerals in the word deletion rules. None of the problems seriously impairs the usefulness of the cloze readability procedure in improving the quality of materials evaluation studies.

Cloze Tests in the Evaluation of the Outcomes of Instruction

Little attention has been given to exploring the potential uses of cloze tests as measures of the knowledge students gain as a result of instruction. The reason may be that educators demand that achievement test questions seem valid, at least intuitively, as measures of the knowledge imparted by instruction. While cloze tests may be made from the instructional materials themselves, it has remained obscure just how a given cloze test item might test the knowledge gained in instruction.

This section will advance the claim that there is a formal similarity between some types of cloze test items and the conventional completion and multiple choice test items generally accepted as tests of the achievement of knowledge. It should be emphasized that the remainder of this discussion is no longer confined to a consideration of just the cloze readability procedure but is extended to the consid-

eration of cloze tests per se---that is, to tests made by deleting any objectively definable language unit according to a set of pre-specified rules.

The argument supporting the claim that there are formal similarities between cloze and many of the conventional achievement test items is based on reasoning that takes this form. Where instruction is given in natural language, it can be analyzed into a list of sentences. Most of the questions that can be asked about a sentence can be expressed as transformations performed on the syntax of the sentence, coupled with the substitution of semantic equivalents for the words and phrases in the sentences. The transformation performed on the syntax of a sentence has the effect of deleting the portion of the sentence which becomes the correct response to the question. The substitution of semantic equivalents can also be expressed as transformation on two or more of the sentences in lists, where the instruction has been systematic. Cloze tests can be produced by these same manipulations.

Conventional Test Items

Instruction as a List of Sentences: Verbal instruction can be usefully regarded as a list of sentences whose truth values have been verified. If one were to construct such a list, it could consist either of the sentences in the exact

order in which they occurred in instruction or of an unordered list that contains a somewhat larger number of sentences. Consider this instructional passage:

> Some arctic explorers were forced to eat raw polar bear meat. Many died before returning home. Polar bears are often infected with trichinosis.

The list of sentences made from this passage may consist of just these sentences in their present order. But, because sentence order in connected discourse transmits information, an unordered list must contain sentences stating the information signaled by the sequence in which sentences occur. There are no well defined procedures for analyzing the syntax of discourse. However, an unordered list made from the instructional passage above might contain sentences 1 through 6 listed below. Presumably, this list of sentences contains all the information contained in each of the sentences taken separately (sentences 1, 2, and 4)

1. Some arctic explorers were forced to eat raw polar bear meat.
2. Many arctic explorers died before returning home.
3. Eating raw polar bear meat has caused the death of some arctic explorers.
4. Polar bears are often infected with trichinosis.
5. The deaths of some arctic explorers has been caused by trichinosis.
6. Some arctic explorers contracted trichinosis as a result of eating raw polar bear meat.

plus all the information contained in the ordering of the sentences relative to each other (sentences 3, 5, and 6).

The act of making an unordered list of sentences may be what instructional programmers loosely refer to as "making explicit" the content of instruction.

Derivation of Sentences: If the instruction in an area of discourse is systematic and complete, the list of sentences either contains or permits the derivation of all (and only) the true sentences that can be stated about that area of discourse. These derived sentences are regarded here as a part of the list, but not as a part of the sentences actually used in instruction.

If the instruction from which the passage above was drawn were systematic and complete, it would have been preceeded by sentences defining the concepts and the relationships among the concepts used in the passages. The following are examples of sentences like some of those that might be found in the preceeding instruction:

7. An explorer is a person who is among the first to examine a region.

8. Polar bears are white bears living in the arctic regions.

9. The arctic is the region lying near the North Pole of the Earth.

10. An uncooked substance is raw.

Ultimately, the instruction would involve contact with concrete objects and sentences naming those objects.

In this discussion, the derivation of true statements about an area of discourse refers to three kinds of behavior.

First, derivation refers to the act of transforming a complex sentence into kernels. For example, sentence 1 can be transformed into the sentences Some explorers were forced to eat meat, The explorers explore the arctic, The meat was raw, The meat was from bears, and The bears lived in the polar region. These kernel sentences can be derived by mechanical processes. Second, derivation refers to the act of deriving sentences such as sentences numbered 3, 5, and 6 which are implied but not explicitly stated in the instructional passage. While most linguists think it likely that sentences of this type may be derivable by a series of relatively mechanical transformations of the sentences used in instruction, they do not presently have sufficient knowledge of inter-sentence syntax to permit us to specify the nature of the transformations for deriving them. Third, derivation refers to the act of substituting for one word or phrase another word or phrase that was equated with it by one of the sentences in the instruction. For example, the sentence Some people who were among the first to examine the region around the North Pole of the Earth were forced to eat uncooked meat from the white bears living in the region around the North Pole of the Earth was obtained largely by substituting equivalent phrases contained in sentences 7, 8, 9, and 10 for the words in sentence 1. Again, note that this is a relatively mechanical process.

Test of Knowledge: The ultimate test of whether a body of knowledge has been mastered is whether the student behaves

appropriately in the environment referred to by the sentences in the instruction. However, practical considerations force educators to settle for less than conclusive proof of mastery, for it is inconvenient to bring elephants into the classroom or to recreate historic disasters, wars, and decisions for the purpose of testing a student's knowledge of things of this sort. Instead, educators rely on some form of verbal response by the student. The student may be asked to write essay exams or to answer objective questions about the instruction.

Having students write essay examinations might be conceptualized in this context as testing a student's ability to select and repeat the sentences actually used in instruction and/or the sentences he derived from the instructional passage. Of greater interest here is the fact that answering objective test items can be conceptualized as filling the blanks left in the sentences.

Question Transformations: Many (and perhaps all) of the verbal questions used in objective tests can be represented as transformations performed on the sentences in an unordered list or on the sentences derived from such a list. An important consequence of this assertion is the fact that a fairly simple set of rules is sufficient to specify the procedures for writing these items and the procedures make the item writing process completely objective and reproducible.

Of broader significance is the fact that this set of rules completely specifies the total population of the test items that can be written over a given unordered list of sentences, making it operationally meaningful to speak of sampling a population of test items over an area of discourse.

There are three general classes of transformations that can be employed for turning a sentence into a question. The first is the yes-no transformation which results in questions answerable by the simple response of yes or no as in the question Were some arctic explorers forced to eat raw polar bear meat. Since questions of this sort are seldom used in testing, they will not be further considered, but much of what will be said subsequently, also applies to the yes-no question. The second is the completion question made by deleting a word, phrase, or clause from a sentence as in Some ---- ----- were forced to eat raw polar bear meat. The third is the wh- question in which a wh- question marker (when, where, how, why, what, how many, etc.) is inserted in the place of a word, phrase, or clause and the word order of the sentence is sometimes rearranged. The question Who were forced to eat raw polar bear meat is an example.

The completion question is perhaps the easiest of all questions to generate. A sentence is selected from the list, a word or phrase is selected from the sentence, and the word or phrase is replaced by a blank space. Table 2 shows the

Table 2

The completion questions that can be made from the sentence in Figure 1

Node	Question
S	Some Arctic explorers (were forced to eat raw polar bear meat).*
S	(Some Arctic explorers) were forced to eat raw polar bear meat.
NP_1	(Some) Arctic explorers were forced to eat raw polar bear meat.
NP_1	Some (Arctic explorers) were forced to eat raw polar bear meat.
MN_1	Some (Arctic) explorers were forced to eat raw polar bear meat.
MN_1	Some Arctic (explorers) were forced to eat raw polar bear meat.
VP_1	Some Arctic explorers (were forced) to eat raw polar bear meat.
VP_1	Some Arctic explorers were (forced) to eat raw polar bear meat.
VP_1	Some Arctic explorers were forced (to eat raw polar bear meat).
NP_2	Some Arctic explorers were forced (to eat) raw polar bear meat.
NP_2	Some Arctic explorers were forced to (eat) raw polar bear meat.
NP_2	Some Arctic explorers were forced to eat (raw polar bear meat).
MN_2	Some Arctic explorers were forced to eat (raw) polar bear meat.
MN_2	Some Arctic explorers were forced to eat raw (polar bear meat).
CN	Some Arctic explorers were forced to eat raw (polar bear) meat.
CN	Some Arctic explorers were forced to eat raw polar bear (meat).
MN_3	Some Arctic explorers were forced to eat raw (polar) bear meat.
MN_3	Some Arctic explorers were forced to eat raw polar (bear) meat.

*Underlined portion of each sentence is the portion of the sentence deleted to form the question.

completion questions that can be formed from the sentence shown in Figure 1. An important feature of the completion question is the fact that, where more than one word is deleted, the deleted words invariably constitute a phrase. This may be verified by tracing the derivations of the deleted words up through the phrase structure tree in Figure 1. The deleted units invariably constitute all the words dominated by a single phrase node. Deletions which cut across phrase boundaries such as The little (boy rode) the horse, virtually never occur in tests. Evidently, the structure of the language requires that all deletions constitute a phrase.

The second feature that should be noted is the fact that structural words (the class of words consisting principally of articles, prepositions, conjunctions, auxiliary and modal verbs, and infinitive markers) are never deleted as single words. In short, questions like (The) little boy rode the horse never occur in tests. When confronted with questions of this sort, people respond by trying to find a lexical word (consisting roughly of verbs, nouns, adjectives, and adverbs) to fit the blank, and complain that the question does not really test their knowledge. Again, this appears to reflect a property of the language.

The wh- question is, in many respects, identical to the completion question. A sentence is selected, a phrase or

word within the sentence is selected, that word or phrase is deleted and replaced with a wh- phrase, and then (usually, but not always) the word order of the sentence is rearranged so that the sentence begins with the wh-phrase. Table 3 shows the wh- questions that can be written over the sentence in Figure 1. The units deleted are either individual lexical words or phrase units, as in the deletion questions.

A variation on the wh- question is sometimes observed in tests. This consists of questions made by replacing a lexical word or phrase with a wh- phrase and then neglecting the step of rearranging the word order. This results in questions like The little what rode the horse or The what rode the horse. Wh- questions of this sort are almost identical to the completion question, the only distinction is the use of a wh- phrase instead of a blank.

By now it should be evident that cloze and the conventional completion and multiple choice items are similiar in that both are made by a deletion process. Moreover, it is possible to define a cloze procedure that would delete only the words, phrases, or clauses that can be deleted by conventional item writing procedures.

Comparison of Cloze and Conventional Tests

Methods of Selecting Test Items: The chief distinction between cloze and conventional tests is in the methods used to select the items to appear in the test. Cloze tests must

Table 3

Wh- questions obtained by applying transformations to the nodes of the sentence structure in Figure 1

Node	Question	Constituent Deleted
S	What happened to some Arctic explorers and Who were forced to eat raw polar bear meat?	were forced to eat raw polar bear meat. Some Arctic explorers
NP_1	Which Arctic explorers were forced to eat raw polar bear meat?	Some
MN_1*	Some of what kind of explorers were forced to eat raw polar bear meat?	Arctic
VP_1	What were some Arctic explorers forced to do?	eat raw polar bear meat
VP_2	(None. This node dominates only one lexical constituent.)	
NP_2	What were some Arctic explorers forced to eat?	raw polar bear meat
Inf	(None. This node dominates only one lexical constituent.)	
MN_2	What kind of polar bear meat were some Arctic explorers forced to eat?	raw
CN	What kind of raw meat were some Arctic explorers forced to eat?	polar bear
MN_3	What kind of raw bear meat were some Arctic explorers forced to eat?	polar

*One occasionally encounters questions like "Some Arctic what were forced to eat raw polar bear meat?" These questions have the effect of deleting the noun. It was omitted here only because it sounds a bit awkward to native speakers of English and because it does not follow the rule of shifting the wh- phrase to the initial position in the sentence.

be made without the intervention of human judgment for the selection of any particular item. Test writers are typically obscure about how they select the particular items they choose to write for an achievement test. When a rationale is offered, it usually involves the writer's or some expert group's subjective feelings about what they call the "important knowledge." This subjective feeling of importance is seldom, if ever, analyzed into a set of objective criteria. Undoubtedly, it includes consideration of the logic of the content area, the social utility of various portions of the content, and some judgment about whether the item is too difficult or easy for the students with whom the test is to be used.

Some rigorous effort is made to develop a taxonomy of instructional objectives. As it is practiced, this is a rather naive gesture, for no algorithm is used for deriving the objectives from the instruction. Hence, the process of selecting the test items to be written usually represents a combination of judgments of what the test maker thinks ought to be taught plus what he anticipates will produce items having good statistical properties. The selection of items to be included in a test from among those items the test maker actually wrote is, when done at all, based upon item difficulty and item inter-correlation indices.

Criterion and Norm Reference Testing: Because of the way in which cloze items are selected, the cloze procedure

seems ideally suited for making criterion referenced tests. While traditional procedures of selecting items may be fairly adequate where normative test information is sought, they leave much to be desired where criterion test information is being sought. First, judgments of what knowledge is important are totally inappropriate when applied at the point of test construction. Judgments of this sort are appropriately made at the points of selecting content, forming the instruction, and interpreting measures of the outcomes. Introducing them into the test construction process creates the possibility that the outcomes of major portions of instruction will be ignored. Specifically, it is the function of criterion measurement to measure whatever is taught---that is, whatever appears in an unordered list of the content of instruction.

Second, traditional methods of measurement provide no objective method of defining the domain of knowledge taught. Since a taxonomy of content is not derived by any specifiable algorithm, there are no criteria by which to judge when it is complete, when it contains content not actually taught, or, for that matter, when it contains two statements of the same content.

The construct of the unordered list of instructional statements is a version of a complete taxonomy which, if it could be constructed for each instructional program, would

quickly to be very costly in terms of scoring time. Another alternative would be to present the item in a multiple choice format, offering among the alternatives the words deleted. The other alternatives might be selected from among the most frequent responses when the test is given in a constructed response format.

By far, the most difficult problem arises from the fact that the ordinal positions of sentences transmit information. For example, The boys get home first followed by They rode horses implies that the riding of horses caused the boys to get home first. When a sentence is taken out of context, the information may be lost. Yet, that information is part of the content. Until knowledge of intersentence syntax is sufficiently advanced, it may be necessary to state this information by subjective methods and place those statements among the sentences sampled.

A final problem is the question of how to remove the effects (or suspected effects) of rote memory from cloze tests. When the student is presented with a cloze test over materials he has never read, his response can hardly be said to have resulted from rote memory; neither can it be said to represent the knowledge he achieved from having read the material. Conversely, when the test is made directly from the materials studied, it is impossible to exclude rote memory as a factor contributing to the responses. This is

a problem with all tests in which the items are derived directly from the materials. Consider the sentence The boful wugs daxed the morf. One can hardly be said to gain knowledge, in any usual sense of the word knowledge, from reading this sentence. Yet most people can answer the questions What kind of wugs daxed the morf and The --- wugs daxed the morf.

This problem may be solved using two measures. First, before any sentence constituents have been selected for deletion, the test maker might go through the sentences and replace randomly chosen constituents with semantically equivalent words, phrases, or clauses. For example, if wugs were defined as durfs who gleb moxes, we could derive the sentence The boful durfs who gleb moxes daxed the morf. The second operation, and this should also be performed before constituents are selected for deletion, is to perform one or more transformations in each sentence so that the sentence retains paraphrase equivalence with the original sentence but no longer has the same syntactic structure. The example sentence might become first The morf was daxed by the boful durfs who gleb moxes and through subsequent transformations it might become the two sentences The durfs who gleb moxes were boful. The morf was daxed by them. The items would then be formed by deleting nodes from the sentences that resulted from these transformations and substitutions.

Summary: A process based on the cloze procedure might be used to make a criterion referenced test over an instructional program in this manner. First, the test writer goes through the text generating sentences that make explicit the information contained in the sequential relationships between sentences. Second, samples of these sentences are drawn to provide tests of whatever number and size seems practically and economically desireable. Third, the test maker randomly selects lexical constituents and substitutes equivalent words or phrases for them. Fourth, the test writer performs one or more syntactic transformations on each sentence in such a way that paraphrase equivalence is preserved. Fifth, he randomly selects from each sentence the node to be deleted and forms the question. (By this time it should be abundantly clear that it makes little difference whether he chooses to write questions in a wh- question format or in a deletion format.) Sixth, he gives the test to a group of subjects in a constructed response format and selects the distractor responses from among the highest frequency incorrect responses. Seventh, he forms the items into a multiple choice format, using the constituent deleted, as the correct response and the highest frequency incorrect responses as the alternatives.

Concluding Remarks

The remarkable thing about the cloze procedure is not

that it produces a new kind of test, for there is actually a formal identity between conventional achievement test items and some of the items that can be made by the cloze procedure. Instead, the unique feature of the cloze procedure is that it presents us with the algorithm for making criterion referenced tests over verbal instructional material. In its every-fifth-word deletion form, cloze procedure provides us with a valid and highly reliable method of measuring the relative difficulties of instructional materials for students. Certainly, this constitutes an important contribution to the evaluation of instructional programs, for once the difficulties of instructional materials have been suitably processed, the cloze procedure provides an appropriate procedure for generating criterion reference test items from those materials.

However, it must be clearly understood that the cloze procedure is not a panacea for the construction of achievement tests. Where the object is to obtain highly efficient norm referenced tests, the cloze procedure is of value only for defining the population of possible items that can be written. Furthermore, all criterion referenced tests developed to measure knowledge gained as a result of studying verbally presented content will be less than ideal for as long as there is no satisfactory way to deal with intersentence syntax.

Bibliography

1. Betts, E.A., Foundations of Reading Instruction. New York: American Book Company, 1946.

2. Bormuth, J.R., "Readability: A New Approach," Reading Research Quarterly, 1 (Spring, 1966a) 79-132.

3. Bormuth, J.R., "Factor Validity of Cloze Tests as Measures of Reading Comprehension Ability," (paper read at the American Educational Research Association in February, 1966b).

4. Bormuth, J.R., "Cloze Test Readability: Criterion Reference Scores," 1966c (in press).

5. Bormuth, J.R., "Comparable Cloze and Multiple Choice Comprehension Test Scores," 1966d (in press).

6. Bormuth, J.R., "Optimum Sample Size and Cloze Test Length in Readability Measurement," Journal of Educational Measurement, 2 (June, 1965a) 111-116.

7. Bormuth, J.R., "Validities of Grammatical and Semantic Classifications of Cloze Test Scores," Proceedings of the International Reading Association, 10 (1965b) 283-286.

8. Bormuth, J.R., Cloze Tests as Measures of Readability. Unpublished doctoral dissertation, Indiana University, 1962.

9. Coleman, E.B., "Developing a Technology of Written Instruction: Some Determiners of the Complexity of Prose," Symposium on Verbal Learning Research and the Technology of Written Instruction, Columbia University, 1966a, unpublished.

10. Coleman, E.B. and G.R. Miller, "A Measure of Information Gained During the Learning of Prose," Texas Western College of the University of Texas, 1966b, unpublished.

11. Davis, F.B., Item Analysis Data, Their Computation, Interpretation, and Use in Test Construction. Harvard Education Papers, No. 2. Cambridge: Harvard University, 1949.

12. Fletcher, J.E., A Study of the Relationships Between Ability to Use Context as an Aid in Reading and Other Verbal Abilities. Unpublished doctoral dissertation. University of Washington, 1959.

13. Friedman, M., The Use of the Cloze Procedure for Improving the Reading Comprehension of Foreign Students at the University of Florida. Unpublished doctoral dissertation, University of Florida, 1964.

14. Gallant, R., An Investigation of the Use of Cloze Tests as a Measure of Readability of Materials for the Primary Grades. Unpublished doctoral dissertation. Indiana University, 1964.

15. Greene, F.P., A Modified Cloze Procedure for Assessing Adult Reading Comprehension. Unpublished doctoral dissertation, The University of Michigan, 1964.

16. Hafner, L.E., "Relationships of Various Measures to the Cloze," In Eric L. Thurston and Lawrence E. Hafner (Eds.), Thirteenth Yearbook of the National Reading Conference, Milwaukee: National Reading Conference, Inc., 1963, 135-145.

17. Harris, A.J., Effective Teaching of Reading. New York: David McKay Company, Inc., 1962.

18. Jenkinson, M.E., Selected Processes and Difficulties in Reading Comprehension. Unpublished doctoral dissertation, University of Chicago, 1957.

19. Kuder, G.F. and M.W. Richardson, "The Theory of the Estimation of Test Reliability" Psychometrika, 2 (1937) 151-160.

20. Lord, Frederic M., "Sampling Fluctuations Resulting from the Sampling of Test Items," Psychometrika, 20 (March, 1955) 1-23.

21. MacGinitie, W.H., "Contextual Constraint in English Prose Paragraphs," Journal of Psychology 51 (January, 1961) 121-130.

22. Rankin, E.J., "Cloze Procedure-A Survey of Research," Yearbook of the South West Reading Conference, 14 (1965), 133-148.

23. Rankin, E.F., Jr., An Evaluation of the Cloze Procedure as a Technique for Measuring Reading Comprehension. Unpublished doctoral dissertation, The University of Michigan, 1957.

24. Ruddell, R.B., An Investigation of the Effect of the Similarity of Oral and Written Patterns of Language Structure on Reading Comprehension. Unpublished doctoral dissertation, Indiana University, 1963.

25. Sukeyori, S., "A Study of Readability Measurement -- Application of Cloze Procedure to Japanese Language," (English Abstract), Japanese Journal of Psychology 28 (August, 1957), 135.

26. Taylor, W.L., " 'Cloze Procedure': A New Tool for Measuring Readability," Journalism Quarterly, (Fall, 1963) 115-432.

27. Taylor, W.L., "Recent Developments in the Use of 'Cloze Procedure'," Journalism Quarterly 33 (Winter, 1956) 42-48.

28. Thorndike, E.L., "Reading and reasoning: A study of mistakes in paragraph reading," Journal of Educational Psychology 8 (1917) 323-332.

29. Weaver, W.W. and A.J. Kingston, "A Factor Analysis of the Cloze Procedure and Other Measures of Reading and Language Ability," The Journal of Communication, 13 (December, 1963) 252-261.