CLOZE PROCEDURE:

A COMPARISON OF EXACT AND ACCEPTABLE SCORING

by

JOHN B. LAING

B.A. The University of Victoria, 1978

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in
THE FACULTY OF GRADUATE STUDIES
(Department of Language Education)

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July, 1988

© John B. Laing, 1988
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of LANGUAGE EDUCATION

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date AUGUST 13, 1988
ABSTRACT

A large number of studies support the use of cloze procedure as a measure of overall language proficiency. Despite the acceptance of the procedure itself there is considerable debate concerning the most appropriate scoring procedure to use. This study set out to compare exact and acceptable scoring of cloze tests in order to determine which is most reliable.

Ninety-two students in two colleges were tested with cloze tests made up of two different passages contained in each test. Pearson product moment correlation coefficients were calculated between exact and acceptable scores and it was found that the two scores correlated very highly at .92 for the whole test scores (made up of two 50 item cloze tests). Correlations for different passages 50 items in length ranged from .83 to .87. These results indicate that the exact method of scoring can be used instead of the acceptable method of scoring with little loss of information. It was found that correlations between different passages ranged from .53 (N=45) to .74 (N=47) indicating that in some cases the scores from cloze tests constructed from different passages may produce different results.
CLOZE PROCEDURE:

A COMPARISON OF EXACT AND ACCEPTABLE SCORING

Table of Contents

CHAPTER ONE: INTRODUCTION 1

Statement of the Problem 1

Background to the Problem 1

Evidence Supporting the use of Cloze tests 3

Defining Cloze Procedure 4

Scoring 4

Definitions 9

CHAPTER TWO: REVIEW OF LITERATURE 11

Introduction 11

The Origin of Cloze Procedure 11

Cloze as a Measure of Language Proficiency 15
CHAPTER THREE: DESIGN

Introduction

CHAPTER FOUR: RESULTS

Introduction

Population

Comparison With TOEFL

CHAPTER FIVE: IMPLICATIONS OF RESULTS

Introduction

Limitations of this Study

Implications of Results

Conclusion

Areas for Further Research
LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIGURE I</td>
<td>Comparison of Exact and Acceptable Scores (Pilot)</td>
<td>49</td>
</tr>
<tr>
<td>FIGURE II</td>
<td>Test Forms</td>
<td>51</td>
</tr>
<tr>
<td>FIGURE III</td>
<td>Pie Chart of First Language</td>
<td>56</td>
</tr>
<tr>
<td>FIGURE IV</td>
<td>Box Plot of Age</td>
<td>57</td>
</tr>
<tr>
<td>FIGURE V</td>
<td>Box Plots for Total Exact and Acceptable Scores</td>
<td>58</td>
</tr>
<tr>
<td>FIGURE VI</td>
<td>Acceptable Scores and Exact Scores with Mean Difference Between Scores Added</td>
<td>61</td>
</tr>
<tr>
<td>FIGURE VII</td>
<td>Scattergrams for TOEFL vs Acceptable and Exact Scores</td>
<td>63</td>
</tr>
</tbody>
</table>
CHAPTER I

Statement of the Problem

In the teaching of ESL and EFL it is necessary to test the proficiency of students in order to measure their ability both prior to and after instruction. It would be ideal if a testing method existed which provided a measure of overall language proficiency. Several studies have established that cloze procedure is an excellent measure of overall language proficiency. However, there is considerable disagreement concerning the scoring and interpretation of cloze tests. It is important that the most reliable method of scoring is used in the measurement of students' overall language proficiency. Therefore, this paper will compare the exact method of scoring cloze tests with the acceptable method in order to determine which is most reliable.

Background to the Problem

Cloze procedure was originally developed in order to measure the interpretability of a message by assessing its redundancy. This was done by deleting words from a written passage, resulting in a passage with a series of blanks throughout the text. The success with which a person replaced the blanks with the words that were originally removed indicated the "readability" of the passage. Cloze procedure has come to be used for many purposes and
the literature concerning its various aspects has become more vast. In addition to its use as a readability measure, Cloze procedure has been used to assess students’ ability to comprehend text and also as a teaching device. One of the major uses of cloze procedure is its use as a measure of language proficiency. Though several studies have shown that cloze procedure is an excellent measure of overall language proficiency there is considerable disagreement concerning the scoring and interpretation of cloze tests.

Cloze tests can be scored in a variety of ways. It will be shown that there are primarily two scoring methods which are of concern to teachers. One counts only words which exactly match the one deleted as correct; the other counts any response as correct which is appropriate in the context surrounding the blank. These two methods of scoring are commonly referred to as exact scoring and acceptable scoring, respectively. This paper will attempt to determine which of these two scoring methods is most reliable. Prior to determining the most reliable of these two scoring methods it will be necessary to show, 1) that cloze tests are worthy of being scored in the first place; in other words, that they appear to be useful tests and 2) that other methods of scoring are not useful for classroom testing.
Evidence Supporting the Use of Cloze Tests

Several writers have suggested that cloze procedure is a useful way of developing tests, stating, for example, that it "offers a method of constructing test materials relatively cheaply and quickly" (Carroll, Carton and Wilds, 1959, p. 4). Cloze tests are said to be convenient and "accurate and reliable" as well (Jonz 1976, page 255). Cloze procedure has been not only described as merely "elegant" (Smith, in Buros 1978, page 1176), John Oller has referred to it as "nothing less than a stroke of raw genius" (1972, p. 106). Other research has also been favorable:

The cloze procedure has been the subject of literally hundreds of research efforts over the past twenty-five years ... the overwhelming preponderance of evidence supports the procedure as an accurate and reliable tool. Though originally conceived to be an alternative measure of readability (Taylor, 1953), it has come to be regarded as a measure of reading comprehension and even as a measure of overall language proficiency. (Jonz, 1976, 255)

In addition to their convenience, cloze tests are said to be a very good measure of skills which closely approximate "real" use of language.

The cloze procedure is one of those tests which, like dictation measures integrative skills, and has much to recommend it. The most important argument in its favor is that it requires the student to perform a task which is not unlike what native speakers do in sending and receiving messages. (Oller et al., 1971, page 187)
There is then, considerable evidence suggesting that cloze tests are very useful measures of overall language proficiency.

Defining Cloze Procedure

Chastain (1979) describes cloze procedure as a "test construction procedure that involves deleting words on some systematic basis and replacing the deletions with blanks which the learner must fill in" (p. 491). In other words, the person developing the test selects a written passage, deletes every fifth word (or seventh, or tenth etc.) and replaces each deleted word with a blank, usually of standard length. The students then attempt to guess the word originally deleted, writing their response in the blank provided. Any response which does not exactly match the original deletion is regarded as incorrect. This describes "standard" cloze procedure as developed by Taylor; there are, of course, several variations possible.

Scoring

One area in which variation occurs is in the scoring of cloze tests. In scoring cloze tests there are basically two options: 1) accept only exact replacements as correct or, 2) accept responses which, though not exact, show understanding on the part of the person taking the test. The second
option - choosing responses which demonstrate comprehension - has been employed in various ways by different researchers. In view of the variety of scoring methods which have been employed, it seems rather an understatement to mention that "various definitions of acceptability are, of course, possible..." (Alderson, in Buros 1978, p. 1173). A further complication is the fact that cloze procedure is not always used for the same purpose. Initially developed as a measure of readability, cloze procedure was then used to measure the proficiency of individuals and eventually as a measure of ESL proficiency. The selection of a scoring method is likely to depend on the purpose of the particular cloze test being used. It is therefore important to emphasize that this study is concerned specifically with the use of cloze procedure in assessing the overall language proficiency of adult students of ESL or EFL. It is not concerned with the use of cloze procedure as a measure of readability, or as a measure of the proficiency of native speakers.

Brown (1980) mentions four different scoring methods; clozentropy, acceptable-response, exact-response and multiple choice. As their names indicate, the exact-word method only allows responses which match the deleted word, while the acceptable-word method also permits responses which "make sense" within the context of the passage (Brown uses the term
contextually acceptable" page 311). Clozentropy is an extremely complicated method of scoring. "In fact, the scoring computations are so complex, computer assistance is practically essential" (Darnell, 1970 p. 44). The complexity of clozentropy makes it impractical for use in most classrooms. Though not as complicated as Clozentropy, multiple choice (MC) cloze can also be removed from consideration as a classroom test. MC cloze is sometimes recommended for its ease of scoring (Hinofotis and Snow, 1980, p.129). However, constructing an MC test is an involved procedure.

A caution to keep in mind is that constructing an MC cloze test is a considerably more complicated procedure than constructing an open-ended cloze test. An MC version requires pretesting with an open-ended task (or some other system) to obtain distractors. It is then necessary to pretest the MC version to check item facility and discrimination. Items which do not discriminate well should be eliminated or modified. (p.133)

Brown also mentions the difficulty involved in developing MC cloze tests. He further notes, however, that in addition to the difficulty involved in its construction, MC cloze can be criticized for its low reliability. Hinofotis and Snow also provide evidence of the lower reliability of MC cloze when compared to "standard" cloze tests (1980 - see tables on page 132 and comments page 133).

Another method of scoring is referred to as "synonymic" scoring (McKenna, 1976). McKenna describes synonymic scoring as allowing "synonyms for the
original words" or words "which were very close to the contextual meaning of those words" (p. 142). This method of scoring seems likely to result in low inter-rater reliability in that it is possible that scorers will disagree in their assessment of the acceptability of particular synonyms. Henk and Sellders (1984), in fact, discovered precisely that. They reported that their data "showed that synonymic ratings varied substantially between raters" (p. 285) and concluded that "this instability renders the synonymic scoring method largely impractical" (p. 286).

It may also be argued that acceptable scoring might be likely to produce unreliable results due to differences of opinion concerning what constitutes an "acceptable" response. There is, however, a significant difference between synonymic scoring and acceptable scoring. While synonymic scores imply some sort of comparison between the originally deleted word and the word supplied as a replacement, acceptable scoring does not. Acceptable scoring allows any replacement which "makes sense" within the context of the passage without any consideration of the original deletion. Synonymic scoring, on the other hand, implies that the replacement word must be similar in meaning or form to the originally deleted word. It is possible that the unreliability reported by Henk and Sellders results from the different opinions
of scorers who found different degrees of "similarity" between deletion and replacement word. Another reason for not using a scoring system which compares deletion to replacement is simply that it is not reasonable practice if the aim is truly to assess language proficiency. Henk and Selders state that "when children supply a word which makes sense, it seems reasonable to give them credit even if it does not exactly match the author's word" (p. 283). Henk and Selders do not follow their own advice. If a word should be accepted because it "makes sense" then it should not be necessary to compare it to the author's original word, which is what synonymic scoring implies. There remain two different scoring methods to be compared; acceptable-response and exact-response scoring. It is essential that the most reliable method of scoring is used for assessment of ESL proficiency. Therefore, the purpose of this study is to compare the exact-word method of scoring with the acceptable-word method in order to determine which is most reliable in testing adult ESL students. In order to provide a background to the problem of scoring cloze tests chapter two will explore the "roots" of cloze and trace the evolution of cloze from a measure of readability to a measure of overall language proficiency. In tracing the development of cloze procedure it will be shown that different writers have used cloze procedure for a variety of
purposes and have made different assumptions concerning the interpretation of cloze scores.

Definitions:

Acceptable response scoring: A method of scoring cloze tests in which responses are counted as correct if they result in a passage regarded as acceptable English by the person scoring the test, even if it does not match the original word deleted.

Cloze procedure: A method of producing tests from written material in which every nth (usually every fifth or seventh) word is deleted, resulting in a passage containing blanks where the original words were deleted. The examinee attempts to fill in the blank with the correct word.

Cloze test: A test resulting from a specific application of cloze procedure using a specific passage, deletion rate and scoring method.

Clozentropy: a complicated form of cloze procedure which involves comparing the scores of examinees with those collected from a reference population. Responses which occur with highest frequency in the reference population are given the most weight.

Construct validity: The extent to which a test is presumed to measure skills regarded as representative of those the tester wishes to evaluate.

Criterion validity: An attempt to show the validity of a test by comparing its results to the scores from a more established test which is referred to as the "criterion".
Exact response scoring: A method of scoring cloze tests which only allows responses which exactly match the word originally deleted.

Face validity: A judgement of the extent to which a test is measuring what it is intended to measure based on superficial assessment rather than empirical evidence.

Language Proficiency: A global term implying both receptive and expressive language abilities.

Multiple choice (MC) cloze: A form of cloze procedure in which the test paper provides alternative responses for the examinee to select from.

Pilot: A preliminary study carried out in order to check procedure prior to the completion of a larger, more comprehensive study.

Speeded test: A test which places a time constraint on the examinees which allows only the most proficient students to complete the test in the time allotted.

Synonymic scoring: A method of scoring cloze tests which requires that the replacement word be similar in meaning or function to the originally deleted word.
CHAPTER TWO

Introduction

This chapter contains discussions of:

a) the development of cloze procedure from a measure of readability to a
   measure of proficiency in ESL,

b) some of the disagreement concerning the acceptability of cloze scores
   as measures of ESL proficiency and

c) the limitations of previous research, particularly involving the reliability of
   cloze scores.

The amount of material written concerning the various applications of cloze
procedure is extensive. It was originally developed as a measure of
readability. It later was used as a measure of the proficiency of the person
writing the test (Taylor, 1957). More recently cloze procedure has been used
as a measure of capability in a foreign language (Heilenman, 1983) and a
method of improving students' comprehension of a particular text (Jongsma
1971; 1980). This paper will be concerned only with the use of cloze
procedure as a measure of ESL proficiency in adults.

The Origin of Cloze Procedure

As mentioned earlier, the "father" of Cloze Procedure is commonly
regarded to be Wilson Taylor (1953). Though Taylor's article was preceded by the work done by Ebbinghaus (1897), "few users of the procedure are aware of the early use of the Ebbinghaus completion method" (Buros, 1978, page 1163). Oiler and Inal (1971) in fact state that "the technique was first employed by W. L. Taylor" (p. 315). Since it appears that most of the present application of cloze procedure derives from the work of Taylor, the discussion of cloze procedure will begin from this point. References to the earlier work of Ebbinghaus are in Buros (1978).

Taylor wrote an article about a "new" procedure for determining the readability of text. Taylor's article appeared in Journalism Quarterly, a journal dedicated not to language instruction but to journalism. He claimed in his first sentence that cloze procedure was "a new psychological tool for measuring the effectiveness of communication" (p. 415). Taylor applied his technique to measuring the "readability" of text. He introduced the term "cloze procedure" to refer to his technique.

At the heart of the procedure is a functional unit of measurement tentatively dubbed a "cloze." It is pronounced like the verb "close" and is derived from "closure." The last term is one gestalt psychology applies to the human tendency to complete a familiar but not-quite finished pattern - to "see" a broken circle as a whole one, for example, by mentally closing up the gaps... (p. 415)

Cloze procedure may be defined as: A method of intercepting a message from a "transmitter" (writer or speaker), mutilating its
language patterns by deleting parts, and so administering it to "receivers" (readers or listeners) that their attempts to make the patterns whole again potentially yield a considerable number of cloze units. (p. 416)

Taylor described how the method worked; pointing out that "the concept of cloze procedure involves both oral and written communication and does not specify any particular kind of "'part' for deletion," (p. 416). Taylor used only written materials and deleted only words. The procedure involved deletion of words from a passage by "some essentially random counting-out system." Each deleted word was then replaced with "a blank of some standard length (so the length would not influence the guessing)." The resulting passage is then distributed to the appropriate population who are asked "to try and fill in all blanks by guessing, from the context of remaining words, what the missing words would be." The number of correct replacements are then totaled and considered as "readability scores" (p. 416). In order to determine the most "readable" passage one simply compares the total scores for the passages chosen. The passage which gives a higher score is the most readable. Taylor, who introduced cloze procedure as a measure of readability, compared two different methods of scoring. In one he allowed only the exact replacement of the deleted word, in the other he gave "half counts" for synonyms. Taylor (1953) found that considering synonyms did not warrant the
extra labor involved as results were "virtually identical to scoring only precise matches" (p. 425).

Taylor found that cloze procedure gave results which "were repeatedly shown to conform with the results of the Flesch and Dale-Chall devices for estimating readability" (p. 411). Taylor felt that cloze procedure could assess factors influencing readability that previous formulae ignored. He pointed out that "element counting" formulas "assume a high correlation between ease of comprehension and the frequency of occurrence of selected kinds of language elements . . ." (p. 417). On the other hand "one can think of cloze procedure as throwing all potential readability influences in a pot, letting them interact, then sampling the result" (p. 417).

One of the factors having the greatest possible effect on cloze test scores was the person reading the passage; a more proficient reader would likely produce higher scores while less proficient readers would have low scores. In fact, measurement of individual test takers was the next step in Taylor's research.

Taylor, having further considered the use of cloze procedure, made a

---

1 The Flesch and Dale-Chall are formulas for calculating the difficulty of text. They are based on the physical features of text, such as the number of common words and sentence complexity.
crucial change in the interpretation of cloze test scores. Assuming that readability also involved not only the difficulty of the text, but also the ability of the reader, Taylor (1957) compared the results of cloze procedure to results obtained from the Armed Forces Qualification Test (AFQT). Taylor found that though cloze and the AFQT were generally similar in the kinds of results they yielded, the AFQT was much more difficult to administer.

Taylor had moved from considering cloze procedure as a measure of readability to interpreting it as an measure of the ability of the person taking the test. The discovery that cloze tests might serve as a measure of proficiency was an important milestone which provided further impetus for cloze research. After all, not only did cloze appear to correlate well with measures of proficiency, it was a very simple test to develop and administer.

**Cloze as a Measure of Language Proficiency**

Soon after Taylor's article, Carroll, Carton and Wilds (1959) completed an extensive study which investigated the possibility that cloze might measure proficiency in a foreign language. In the first major paper investigating the use of cloze procedure as a measure of achievement in foreign languages, Carroll et al. concentrated "upon the improvement of tests concerned with the written aspect of a foreign language" (p. 1). Their goal was "to supplement or replace
the kinds of test items found in the written College Board achievement examinations in foreign languages" (p. 1). After concluding their study, one of the most comprehensive done on cloze procedure, Carroll et al. concluded that:

The word-cloze and letter-cloze procedures developed here may be suitable testing devices to assess group differences in second language competence, but they are inadequate as measures of individuals because they are relatively unreliable and are too heavily affected by various sources of extraneous variance. (p. 116)

Carroll et al. mention several possible reasons for variation in the results of cloze procedure; one of the more mundane being simply guessing. Another problem is more difficult to come to terms with.

Another difficulty is the very likely possibility that difference in "intelligence" or other extraneous dimensions of human variation might mask the variation in foreign language competence which this type of test seeks to measure. If such were found to be the case, it would be necessary to consider the advisability of trying to adjust for the influence of this disturbing variable. (p. 3)

This excerpt is very interesting in light of an ongoing controversy concerned with discovering exactly what language proficiency is. To some, "language proficiency" and "intelligence" may be related.

More recently, I have asked whether deep language skills may themselves be identical with what was formerly called "intelligence"... Is it possible that the pragmatic mapping of experimental context into abstract propositional forms might be the characteristic function, even the essence of intelligence itself (Oller 1983: p. 1)
In other words, what Carroll et al. refer to as a "disturbing variable" may be, according to others, very much a part of the construct "language proficiency."

Another possible source of spurious variance mentioned is the effect of the text itself. Carroll et al. correctly point out that "the individual's ability to restore a text depends partly upon the nature and difficulty of the text itself (p. 4). They also found that two passages were "excessively difficult" and thus did not include them in final calculations.

A related problem considered by Carroll et al. was the influence of the method of deletion chosen. The pattern of deletion determines not only the sample of deleted words, but the remaining context. A deletion rate of every fifth word will not only delete different words than an every tenth word deletion rate, it will also leave less words between blanks to provide contextual clues. Carroll et al. state that since they wished to provide the "examinee abundant evidence" (p. 21) and also since they "had no preconception" (p. 22) as to what classes of words would be most effective, they chose an every tenth deletion pattern. Deletion was adjusted to avoid proper names, numbers or numerical expressions (p. 23) and hyphenated words were counted as two words (p. 22). It was found that "different deletion versions can and do sometimes produce significantly different scores for the passages" (p. 58). It
should be noted however, that the experimental design in this case is flawed. The comparison of the different deletion patterns also involved different examinees. Though Carroll et al. attempted to keep the "ability to do cloze items constant" by using scores from a third passage which both groups did in common" (p. 58), it is possible that matching did not successfully account for other factors. As Campbell and Stanly (1966) point out, groups should be equated through randomization rather than matching. They refer to matching as "intuitively appealing and misleading" (p. 2). Moser and Kalton (1971, p. 82) state "to ensure true randomness, the method of selection must be independent of human judgement." (quoted by Johnson in Read, 1981, p. 179).

The first major work by Carroll et al. involving the use of cloze procedure as a measure of second language proficiency produced rather disappointing results, at least for those who hoped cloze procedure represented a useful new assessment.

According to Stansfield (1977) "a full decade passed before another published study directed attention to the cloze procedure in the assessment of language proficiency" (p. 3). Stansfield's statement ignores several articles such as a study of cloze procedure as a
measure of reading proficiency (Rankin, 1959) and a factor analysis of a test battery including several cloze tests (Weaver and Kingston, 1963). It is true, however, that writers did not seem to be interested in cloze procedure as a measure of second language proficiency. Stansfield points out that this lack of interest may have been due, at least in part, to the critical comments of Carroll et al. coupled with cloze procedure's "lack of face validity" (1977, p. 3). Weaver and Kingston also mention the low face validity of cloze.

The next flurry of interest in cloze procedure as a measure of second language proficiency seems to have occurred roughly around 1970. Oller and Conrad (1971) list several articles which followed Carroll et al. and "advocated use of the cloze procedure for ESL either as a teaching or testing device" (1971, p.184). The paper that seems most frequently cited is that of Darnell (1968; see also Darnell 1970)

**Clozentropy**

Darnell used a rather complicated method of scoring cloze tests. He felt his method was enough of a breakthrough to be referred to as "a new test of language proficiency" (1968, p. 2). His goal was to develop "procedures which permit an alternative to the right-wrong scoring
system and to the subjective evaluation of essays” (p. 4). This was to be accomplished by weighting each person's responses against the responses of a criterion group, a procedure anticipated, at least in concept, by Carroll et al.

On a priori grounds, however, a case would be made for scoring in terms of community of response, i.e. giving a positive weight to any response of high frequency in a formative sample. Such a scoring procedure might enhance reliability and even validity because the "correct" response would correspond to a sort of linguistic norm rather than to the possible idiosyncratic item found in the original text. (p. 23)

In order to score relative to a "community of response," Darnell combined cloze procedure with an "entropy measure derived from information theory which indexes the compatibility of an individual's responses with those of a selected criterion group" (1968, p. 2), thus the term "clozentropy".

Pilot studies revealed that clozentropy only correlated at .62 (p < .01) with "standard" cloze procedure which scored only exact replacements as correct. On the basis of these results, Darnell concluded that Clozentropy gave different results than cloze; results which might perhaps be more "favorable" (p. 111).

A battery of four Clozentropy test passages was administered to 48
foreign students in conjunction with the Test of English as a Foreign Language (TOEFL). A Hoyt reliability coefficient was computed both for the total Clozentropy and the TOEFL. Both coefficients rounded to .86 and were regarded to be "satisfactorily high" by Darnell. It was also found that Clozentropy correlated highly with the listening subsection of the TOEFL and neither TOEFL or Clozentropy correlated well with academic achievement. On the basis of his findings, Darnell concluded by noting advantages which Clozentropy has over "the more common kind of tests:"

1) Ease of administration

2) Ease of creating additional forms

3) Alternative forms make cheating difficult.

4) Allows a free response rather than merely selection from a list of alternatives (pp. 31-32)

As Darnell himself points out, a major limitation of clozentropy is its complex scoring method; a feature which effectively prevents its widespread use in classrooms.

One major limitation of the CLOZENTROPY test is the scoring procedure. It would be practically impossible to tabulate the responses and compute the . . . scores for any reasonable number of subjects without computer assistance. (1968, p. 32)

Though the technique developed by Darnell is complex, it has been
mentioned here briefly because it may be of interest to more technically minded researchers. Also, Reilly (1971) has written a paper simplifying the derivation of scores according to Darnell's method. The simplification given by Reilly may not be regarded as "simple" by many teachers. His method still requires a criterion group for establishing a set of scores to be used as a "standard." It is thus necessary to administer at least two cloze tests, one to the people being tested, and the other to the people serving as the criterion group. Further, the weight of each criterion response must be calculated for each blank.

Miller and Coleman (1967), concluded that scoring by the exact response method was the most effective method since alternative methods did not produce substantially different scores. However, in applying cloze procedure to an assessment of language proficiency, the work of both Miller and Coleman and of Taylor should be regarded with caution for three reasons. First, only giving part scores for synonyms weights the score in favor of exact responses. "This has the effect of biasing the results in favor of the exact word procedure" (Alderson, 1978, p. 1171). The second reason the scoring issue has not been resolved is that readability was being studied rather than student proficiency. Since readability studies average the scores of many students in
order to "place" a text on a scale of difficulty, the variation caused by different scoring procedures may be masked by the averaging of scores. Finally, though it was noted that perhaps scoring of cloze tests would account for each "good enough" response, responses were considered "good enough" only if they reflected the meaning or word class of the original deletions; a practice which does not seem appropriate for measuring proficiency rather than readability. Sampson and Briggs (1983) point out that "this procedure does not take into account words which are contextually appropriate but are neither exact replacements nor synonyms" (p.178). Whether or not a response is "good enough" should be determined without referring to the original deletion. It should also be noted that both Taylor and Miller and Coleman tested native speakers. It is possible that different results would have been obtained with students learning English as a second language.

Following the work of Darnell there was an increase in the attention directed toward cloze procedure as a measure of "overall second language proficiency" (Stansfield 1977), though the research dealing with non-native speakers remained sparse. Somewhat more recently, Alderson noted:

With non-native speakers, not a great deal of research has been done, but what there is suggests cloze correlates well with measures of EFL proficiency. See, for example, Oller and Conrad (1971). Oller (1973a), Irvine, Atai and Oller (1974), Stubbs and Tucker (1974), Aitken (1977), Streiff (1978). (Alderson 1983, p. 206)
The study by Oller and Conrad (1971) mentioned by Alderson differed from previous research in that cloze procedure was selected as part of an "integrative" rather than a "discrete point" approach to testing.

In listening, we anticipate what the speaker will say next and frequently (either overtly or covertly) supply missing words or phrases. In speaking, we sometimes find ourselves groping for a word halfway into a sentence. These processes have their counterparts in reading and in writing (Oller, 1971b). For instance, all accomplished readers sometimes creatively supply words or even whole phrases not in the text (Goodman, 1969). (p. 187)

This excerpt reveals what Oller in particular has come to believe cloze tests measure. He has come to refer to this overall competence as a "grammar of expectancy" (Oller, 1973, p. 113).

Oller and Conrad deleted every 7th word from a 350 word passage leaving "the first few lines" intact (p. 188). The cloze test was given to foreign students entering UCLA. The students had previously been divided into five proficiency levels. It was found that the cloze test ranked the students as expected with one exception (see following paragraph). The results from the cloze test were also compared to the results from Form 2C of the UCLA ESL Placement Examination which 35 students had taken previously. "The coefficient of determination was .77 indicating that 77% of the variance in the cloze test was present in the UCLA, ESLPE, Form 2C). The highest product-moment
correlation correlations were obtained between cloze with Dictation (.82) and cloze with reading Comprehension (.8)" (p. 191).

Oiler and Conrad were surprised to discover that there was a difference in scores between English as a Native Language (ENL) college freshmen and ENL graduate students. On the other hand, there was no spread between ESL students and ENL college freshmen. This finding lead Oiler and Conrad to comment on the results of another study (Oiler 1971) which "demonstrated that the differentiation of ESL speakers is significantly superior when a different scoring method is used - when any contextually acceptable word is counted as correct (Oiler and Conrad 1971, p. 191).

Oiler and Conrad felt that their results gave support to the finding of Darnell that cloze correlated best of all with dictation (.82) (p. 192). They also mentioned that the highest correlations were between cloze and what they regarded to be "measures of integrative skills." Their results lead them to the optimistic conclusion that "with further experimentation and refinement, the cloze method may play an extremely useful role in the placement of non-native speakers of English and in the diagnosis of their special language problems" (p. 192).

Oller (1972) later went on to do further research with cloze procedure. He
again mentions the idea that cloze may perhaps provide an excellent measure of overall language skill. He assumes that "the foundation of all language skills is the capacity to anticipate elements in sequence" (p. 151). Based on this assumption, Oller favored cloze procedure as a device suitable for the measurement of skills linked to an underlying overall language proficiency.

It is this little understood element of expectancy which seems crucial to the successful production and interpretation of utterances. Since it seems likely that the cloze procedure - a method of test construction which consists of deleting words from prose is apt to be an excellent device for testing this sort of expectancy, I predict that it will play an increasingly important role not only in the measurement of second-language proficiency, but also in experimental exploration into the mysteries of the human mind itself. (p. 151)

Despite his enthusiasm for cloze procedure, Oller felt that questions remained unanswered. The first issue he raises is that of scoring. He points out that exact word scoring may be inappropriate because it creates a test too difficult for non-natives and also because it is "intuitively unsettling." In other words, there are cases when a response is wrong only because it does not match the original deletion.

Suppose, for example that an item reads "the ______ went down to the stream." If the exact-word is, say "child," is it reasonable to class "horse," "dog," "animal," etc., along with clearly incorrect fill-ins like "of," "and" "table," etc.? The task of guessing the exact-word is not necessarily a language skill in the ordinary sense of
In order to determine the usefulness of allowing contextually acceptable responses, Oiler deleted every 7th word from three passages roughly 373 words in length. The passages ranged in difficulty from "beginning ESL" to "advanced ESL."

All subjects also completed the UCLA ESL Placement Examination Form 2A Revised. Oiler not only discovered that acceptable scoring correlated highest with the UCLA ESL PE he also found that acceptable scoring "is superior in terms of item discrimination and validating correlations regardless of the level of difficult of the test" (p. 157). Further, as he also concluded in an earlier study with Conrad (1971), Oiler again found that "cloze tests tend to correlate best with tests that require high level integrative skills" (p. 157).

Enough evidence had accumulated to support the use of cloze procedure that Irvine Atai and Oiler (1974), used it as a standard in order to investigate the Test of English as a Foreign Language (TOEFL). Irvine et al., set out to use "dictation and cloze procedure as a criterion against which to compare TOEFL scores " (p. 246) .

The TOEFL was administered to 159 native speakers of Farsi. They were asked to return at a later date to complete cloze and dictation tests. Two
passages were recorded on tape for the dictation. The cloze test consisted of a passage of 394 words with every seventh word deleted. The cloze test was scored by both the exact and acceptable methods. Irvine et al, note that the correlation between the exact method and the acceptable method was sufficiently high (.94) to allow the use of the exact method, particularly in situations where the person scoring the test is a non-native speaker.

Irvine et al. found that cloze correlated best with the dictation subsection of the TOEFL with the exception of the listening comprehension (LC) subsection. The cloze correlated better with the combined dictation (.75) than with any TOEFL part score except LC (.76). It correlated at .79 with the TOEFL total score. It is interesting that the .75 correlation between cloze and dictation total was higher than the correlation of any TOEFL sub-section with any other subsection. (p. 250)

On the basis of these results, Irvine et al, concluded that what dictation, cloze and the listening comprehension section of the TOEFL measure "taps a source necessarily common to both writing and speaking skills; namely, the learners' underlying language competence or internalized expectancy grammar" (p. 250).

The authors concluded that on the basis of correlations with cloze tests
and dictation, the listening comprehension section of the TOEFL "probably shows the best estimate of the efficiency of the internalized grammar of the non-native speaker" (p. 251).

Though Irvine et al. suggest that exact replacement scoring can be used as a replacement for acceptable response scoring on the basis of high correlation between the two, they also mention that "acceptable-cloze correlated more highly with the dictation total (.75) than did exact-cloze (.69) (p. 250).

Krzyzanowski (1976) attempted to "establish how a cloze test measures the language proficiency of a non-native learner" (p. 29). He administered two test batteries, each of which included a cloze test, a vocabulary test, a dictation and a structure test. The cloze test in each test battery was taken from the same passage. The test batteries were given to three groups of students. One group (referred to as group III) was somewhat more proficient than the other two, as those students had been given a "special program of English" (p. 30). The contextually acceptable" scoring method was used on the basis of previous research (already mentioned) done by Oller (1972), though he notes "in a number of cases it was extremely difficult to decide whether or not a word is contextually acceptable" (p. 31). It was found that in the first test battery the
cloze portion of the test gave results similar to dictation; results which again confirmed the previous work of Oiler. In the case of the most proficient group of students it was found that scores had little variance and thus did not discriminate as well as with the other two groups.

The discriminatory power of the test tends to increase with the increase of the test difficulty . . . In other words, this means that this test measured the language proficiency at the intermediate level better than at the advanced level (though it was efficient at both). (Krzyzanowski, 1976, p. 32)

This excerpt indicates that an important relation may exist between the ability of the test taker and the results of cloze tests; possibly an important factor to consider during the selection of passages for cloze test development.

Krzyzanowski's results differed from Oller and Conrad's (1971) in that the cloze test did not correlate best with dictation except for the most proficient group. He states that "it may be that this assumption is true only for more advanced learners" (p. 33). In contrast to Oller and Conrad, the highest correlations were with the vocabulary section of the test battery.

To investigate "how some changes in the structure of a cloze test influence its results" (p. 36), four versions of the same cloze test were created. The second test battery was given to four groups of students who were chosen in order to make them as comparable as possible in terms of language proficiency. In CLOZE IIA every seventh word was deleted. In CLOZE IIB
deletions were made in order to alter the ratio of content to function words. In CLOZE IIC every sixth word was randomly deleted. Finally, CLOZE IID had every fifth word deleted. It was found not only that the different versions of the cloze test differed in difficulty, but that the rate of deletion was not a good indication of difficulty. It is also necessary to consider the nature of the words deleted. Krzyzanowski discovered that the most difficult test of the four versions of the second cloze test deleted the highest proportion of content words.

The more content words are deleted, the more cues to the content are missing and the less the whole text becomes comprehensible. (p. 39)

However, Krzyznowski notes that "the discrimination power of the cloze items does not depend on any formal criterion external to the text" (p. 42). The main factor contributing to item difficulty is, according to Krzyzanowski "the immediate context of the deleted words."

Finally, it is concluded that a cloze test is a good assessment of language proficiency particularly if it is included as part of a battery of tests. Krzyzanowski states that since cloze items depend "on the immediate context of deletions, its applicability as an achievement test seems very limited." (p. 43). However, this last statement is in direct contrast to his earlier conclusion that cloze tests can be a "very good instrument" for assessment. It is not clear
what is meant by "achievement." Krzyzanowski also concludes, through an
analysis of student's errors, that cloze tests measure different "elements of
language skills at different proficiency levels" (p. 43). He notes that it is
possible to manipulate the items of a cloze to alter the difficulty of the test.
However, according to Kryzysanowski, the various types of deletion produce
unpredictable results. Such unpredictability would tend to contradict any
attempt to manipulate items, since the results of such manipulations would
presumably be similarly unpredictable. Krzyzanowski does not provide a
procedure for developing cloze tests which would produce tests of a particular
level of difficulty.

An important point demonstrated by Krzyzanowski is that various deletions
may have an influence on the overall results of cloze tests. Further, the
proficiency of the subjects may also affect results.

In 1977, nearly twenty years after Carroll et al. completed their study, it
seemed as though more teachers were becoming aware of cloze procedure.

The "cloze" test has no doubt become one of the most talked about

Based on the results of previous writers (Darnell 1968; Oller 1972; Stubbs
and Tucker, 1974), Aitken encouraged teachers to use cloze procedure:
The purpose of this paper is to provide some guidelines on the construction, administration, scoring, and interpretation of cloze tests, as well as to discuss a possible explanation of the cognitive processes involved in taking a cloze test. It is hoped that this discussion of the cloze procedure will be comprehensive enough to allow ESL teachers to use the cloze procedure in their own classrooms. (1977, p. 54)

Aitken advises the reader that constructing a cloze test is "simple." He favors the "any-contextually-acceptable-answer method of scoring, though he acknowledges the need for more research on the subject of scoring. Incorrect spellings were accepted as long as the word was "recognizable" and "grammatically correct."

Aitken thus attempts to move cloze tests from the realm of speculative research to the classroom. He concludes his paper by endorsing cloze procedure as an effective method of evaluation.

In conclusion, I have found, after having constructed, administered and scored over a thousand cloze tests to ESL students in the last three years, that the cloze procedure is an extremely simple, yet valid language proficiency test. Cloze procedure is not a panacea for all ESL testing problems; however cloze tests yield more "miles per gallon" of sweat spent in test construction than most ESL teachers realize. Overall language proficiency tests like the cloze merit much greater consideration than they are given in preparing ESL proficiency test batteries. (p. 66)

At this point, it seems that cloze procedure is a very reasonable way of testing ESL proficiency. However, not every review of cloze procedure concludes by endorsing it.
Unfavorable Reviews of Cloze Procedure

In contrast to Aitken's enthusiasm, Porter (1978) feels that there is reason for caution. He notes that "tests using cloze procedure are growing increasingly popular" yet several questions remain unanswered.

People are rushing to make use of cloze procedure as a weapon in the teacher's armory, without pausing to ask the fundamental questions: 1) what is the test for, what does it do? and is the weapon worth having - that is, does it do a consistently good job? (p. 333)

In order to answer these questions, Porter set out to provide information to the following, more specific problems:

1) Will two cloze tests constructed from a single passage, but with deletions beginning at different points be equivalent?

2) Will students be ranked similarly by passages written in "different styles?"

3) Will the "contextually acceptable method give better differentiation of students that the "exact word" method?"

Porter constructed two cloze tests. Both tests were identical with the exception that in the second cloze test, each deletion was one word earlier. Each test was divided into 2 subtests of 50 blanks: one literary, the other non-literary. These two passages were chosen to see if writing style would
produce some effect. Test 2 followed test 1 by an interval of 3 weeks. Porter dismissed any practice effect as "irrelevant." (p. 336).

Analyzing his results, Porter found that, according to the low Pearson Product Moment Coefficient of correlation between the two tests, (.57 for exact scores; .65 for acceptable scores) the two tests were not equivalent. Porter also found that the subtests produced different results, indicating that style is possibly an important influence on cloze test scores (Porter notes that this result may be due to deletion pattern rather than style). He found that acceptable word scoring was generally better than exact replacement scoring in discriminating between subjects, but not between year-groups (p. 340).

In contrast to most of the writers before him, Porter not only does not endorse cloze procedure, but feels that it should not be used by classroom teachers. He thus expresses a view of cloze in complete opposition to that of Aitken.

One general conclusion, however, would seem inescapable: cloze tests should not be constructed indiscriminately by the classroom teacher - or indeed anyone else - but only on an individual basis by professional test-constructors who are in a position to check correlation between the results yielded and those obtained on independently validated tests. But if this must be done, the attractiveness of the technique is largely lost. (p. 340)

Clearly, if Porter's results are to be taken seriously, there is cause for concern regarding the use of cloze procedure in evaluation.
Alderson, (in Buros, 1978) around the same time as Porter, came to a
similar conclusion in his review of the research on cloze procedure. He points
out that, because of differences in deletion rate, passage, scoring etc, cloze
tests may not be comparable.

The problem of the noncomparability of different cloze tests -
based on different texts, deleting different words, at different
frequencies, and scoring responses differently- bedevils an
evaluation of the cloze procedure, since what may well be true for
one application of the procedure to one test may be invalid for
another application. It is an apparently simple tool for the test user
since all one has to do is select a text, remove some words, and
see how close to the original text the subjects come, with no
further knowledge of test construction or interpretation needed. But
it is in this very simplicity that the danger lies since there can be no
guarantee that one has produced a good test as a result of the
application of the procedure. . . This is probably enough to make
any test user very wary in his use of such a procedure without
further investigation. (p. 1174)

Alderson (in Oller, 1983) completed a study which investigated the effect
of three different tests, different deletion rates and five different scoring
procedures. Results were compared with several external measures.
Alderson found that different formats seriously influenced the results of cloze
tests.

The major finding seems to relate to the deletion rate variable, in
that changing the deletion frequency of the test produces a
different test which appears to measure different abilities
unpredictably. (p. 211)
Alderson (1983) completed another study which investigated the effect of three different tests, different deletion rates and five different scoring procedures. Results were compared with several external measures. Alderson found that different formats seriously influenced the results of cloze tests.

The major finding seems to relate to the deletion rate variable, in that changing the deletion frequency of the test produces a different test which appears to measure different abilities unpredictably. (Alderson 1983, p. 211)

Cloze procedure, according to Alderson and Porter, must be used with caution. It is apparent then, that though several writers vouch enthusiastically for cloze there is reason to use it with care.

It has been shown that considerable evidence weighs in favor of using cloze procedure. Jonz was earlier quoted (page 3 this paper) as saying the "overwhelming preponderance of evidence supports the procedure as an accurate and reliable tool." In fact, if there was nothing to support the use of cloze procedure, there would be little point in comparing methods of scoring cloze tests. On the other hand, a few writers have pointed out several reasons for interpreting the results of cloze tests with a "grain of salt." The particular volume of salt recommended varies with different authors, Oller advocates cloze results be ingested "salt-free", while Alderson would probably
recommend a pound or two.

Given the number of writers in favor of cloze procedure there is clearly evidence of its usefulness as a method of producing tests. On the other hand others suggest (Porter 1978, Anderson 1978 and Klein-Braley in Oller 1983 for example) that cloze procedure should be used with some degree of caution or not at all. It may be that both conclusions regarding cloze procedure may be correct. The reason for such strong differences of opinion concerning cloze tests may stem from the nature of cloze procedure itself. It is important to distinguish between cloze procedure, which is a method of developing tests, and any particular cloze test, which is one specific application of cloze procedure utilizing a specific passage, deletion rate and scoring method. Cloze procedure is frequently thought to work successfully with virtually any text, an assumption usually attributed to Oller.

In spite of the natural feeling that just any old text will not do, or that a suitable text must be a carefully chosen and edited text, research has shown that the cloze procedure is probably appropriate to just about any text. True, some may be more difficult than others, but it has been demonstrated that for some purposes (e.g. testing ESL proficiency of university level foreign students) the level of difficulty of the task does not greatly affect the spread of scores that will be produced. (Oller, 1979, p. 364)

There is a tendency in the research performed on cloze tests to assume what I would call the notion of cloze equivalence across tests, the idea that any cloze test is equivalent to any other cloze test, the assumption that a cloze is a cloze is a cloze. (Klein-Braley
It is possible that the cloze procedure may result in individual tests which differ in characteristics. In fact, the assumption that cloze procedure may utilize any passage without influencing results is, as Oller himself notes, counterintuitive. Assuming the possibility that cloze tests may differ, it is also possible that cloze procedure per se may provide an excellent method of evaluating ESL and EFL students' overall proficiency, even if particular cloze tests do not. Johnson points out that there is no conflict in the finding that while most cloze studies support the use of cloze procedure, other findings have been disappointing.

However, there is no necessary conflict between the results obtained by Stubbs and Tucker and Alderson, which suggest the eminently reasonable conclusion that the effect of changes in deletion rate will vary, depending upon the characteristics of text and subject. (Johnson in Read, 1981, p. 178)

Recapitulation

Having examined the development of cloze procedure a few key points become clear. Possibly the most obvious is that the "bibliography on cloze is vast" (Alderson, 1983, p. 206). It then becomes clear that most of the writers support the use of cloze procedure as a measure of overall language proficiency.
It produces high reliability per unit of time, and is generally popular with pupils. (Elley, in Burrows 1987, p. 1176)

... a quick, economical method of measuring overall language proficiency. (Brown, 1980, p. 311)

... cloze tests tend to correlate best with tests that require high-level integrative skills. (Oller, 1972, p. 157)

There is now a considerable body of research providing sound evidence for the predictive validity of cloze test scores. (Bachman, 1982, p. 61)

However, it is also obvious from reviewing the literature concerning cloze procedure that several writers disagree with the widespread view that cloze procedure is a useful test of ESL proficiency.

There are a number of myths quite widely held about cloze testing, which have arisen perhaps out of wish-fulfillment rather than the literature itself, and are based on the false premise that cloze is "an automatically valid procedure which results in universally valid tests of language and reading" (Alderson, 1979, p. 220). (Johnson, in Read, 1981, p. 177)

Given these observations, it is clear that a further study of cloze procedure is needed, particularly concerning scoring, since it appears that exact scores may produce different results from acceptable scores. How should a study be conducted in order to determine the best way to score cloze tests? Fortunately, a review of the material concerning cloze has revealed several possibilities for improving the design of studies concerned with cloze procedure. In uncovering areas of possible improvement for cloze research,
Klein-Braley (in Oller, 1983) has perhaps been of most help in stating that
cloze research usually lacks 1) a large population of results and 2) repeated
measures using the same population of test takers.

In particular we need evidence collected from two (or more cloze
tests administered to the same examinees. (Klein Braley, p. 222)

Another weakness of previous research is the fact that several writers have
failed to randomize treatments. In other words, rather than administering two
different types of test randomly to examinees, each form was distributed to a
particular group. This means that any difference in results may be due to
differences in examinees (see for example Carroll et al. 1959 and Porter
1978, pp. 335-336)

There seems to be some support among authors for the acceptable
method of scoring cloze procedure, though there is some confusion as to
what is meant by "acceptable." Alderson used five different scoring
procedures and found that "scoring for any semantically acceptable word
(SEMAC) produces among the highest correlations with the ELBA total"
(1983, p. 208). Alderson describes the ELBA as a test of proficiency in English
as a foreign language." (p. 207). Porter, though not endorsing cloze
procedure, does note that "there is a tendency for acceptable-word scoring to

41
differentiate better between subjects for whom English is a foreign language (Porter, 1978, p. 338)

Johnson goes on to suggest, however, that it may be difficult to determine what is "acceptable." Previously, Krzyzanowski was quoted as finding it difficult to determine the acceptability of some responses (page 30 this paper). No author cited has given a very concise definition of scoring criteria. Brown mentions that "the AC method usually counts any contextually acceptable answer as correct" (1980, p. 311). Alderson used "any semantically acceptable word" as a criterion without further clarification. (1983 pp 207-208).

It is possible for a response to be quite different semantically from the original deletion and yet still result in a sentence which "makes sense." Of course it is difficult to define "acceptable" in precise terms because so many variables are involved. This difficulty must be taken into account in any analysis of scoring.

Many people have served as subjects in cloze testing research. Generally these subjects are quite proficient speakers of English. There appears to have been little effort made to obtain a more general sample of the population. There are three main reasons for this. The first is that most of the research done on cloze procedure has been done by people in universities. Such researchers tend to use the most available subjects, which are students at
their universities. The second reason for not including less proficient students is simply that there is a minimum level of proficiency required to complete a cloze test. In order to complete a cloze test, a student must at least be able to read. The third reason is simply that researchers are not always interested in less proficient students. Carroll et al. (1959) were interested in seeing if cloze procedure could be used to "supplement or replace foreign language tests at the college level. Similarly, one of the most prolific writers on cloze, John Oller has been concerned mainly with the foreign student population at UCLA. This concentration has several drawbacks. Not only is the university population fairly proficient in general (very poor students are usually not admitted) but the situation in which they have been taught may be similar. More specifically, the oral and written language are more likely to have been taught in tandem. In situations where grammar-translation is often the mode of instruction (such as in Japan), and the spoken language is not emphasized, the results of cloze tests may require different interpretation. The high correlation between cloze scores and "integrative" parts of the TOEFL such as listening comprehension referred to by Oller may simply be the result of a particular method of teaching. In situations where students are exposed mainly to written work, cloze scores may give an inflated idea of overall
Kenji Ohtomo (1981) administered a battery of tests including a cloze test to Japanese university students.

It is interesting to note that the results of the cloze test, which is concerned exclusively with written language, should correlate so highly with spoken language test results. (p. 472)

Ohtomo's comment is somewhat misleading since the correlation he regards as "high" would not be accepted as high by most others. In one experiment the correlations between cloze and listening comprehension and dictation were .5 and .4 respectively. In a second experiment the highest correlation was between cloze and dictation at .62. It is important to note that Darnell regarded a correlation of .62 between tests as sufficiently low to indicate that they were not measuring the same thing. In his article Ohtomo also mentions another study completed by Lewis Pike (1976) in which three groups of different nationalities were tested with cloze and the TOEFL. It was found that for all groups the listening comprehension portion of the TOEFL "ranked lowest or next to the lowest in correlation with the standard cloze test" (Ohtomo, p. 473).

The results mentioned by Ohtomo conflict with those obtained both by Darnell (1968) and Irvine et al. (1974), leading Ohtomo to conclude that for Japanese subjects it may be not be appropriate to use cloze tests to predict proficiency.
listening comprehension. When cloze tests are given to students who differ from the "usual" population studied in cloze research, results may differ also.

The Effect of Deletion Pattern and Text

The pattern of deletion can be thought of as the way in which deletions are made. There is, of course, a relation between text and the effect of particular deletions. The removal of a word simultaneously determines the item to be replaced as well as surrounding contextual clues. Taylor attempted to make his method as simple as possible, "so simple that anyone who can count to ten can do it for any sort of material, regardless of its topic or difficulty." (p. 418). Since Taylor was interested in using several scores to "rate" the difficulty of a passage he was not concerned with the inevitable slight variations that might exist from test to test. Later, when cloze scores became used as a measure of proficiency, - the possibility of fluctuations due to differences among cloze tests became a much more serious problem. Generally, deletions have ranged from five to ten words for reasons summed up by Pennock (1972).

Bormuth (1968) has recommended the deletion of every fifth word, leaving an underlined blank of fifteen letter spaces " when validity, economy and convenience are considered simultaneously." This advice follows the findings of MacGintie (1961) that every fifth word deletions were independent, and Taylor's (1956) conclusion that such deletions are as far apart as they need to be. Every fifth word of the passage should therefore be deleted . . . this ratio
provides a sufficient number of items to enable reliability, while at the same time keeping the test to a practical length. Fifty deletions on a passage of approximately 250 words is recommended. (p. 36)

Oller (1973) notes that "typically every fifth, sixth, or seventh word is deleted from a written or spoken passage" (p. 107). None of these authors mentions any need to be concerned about the effects of different deletions.

Klein-Braley (in Oller, 1983) states that one of the fundamental assumptions of cloze tests is that a "mechanical" nth word deletion produces a random sample of all elements of the language" which is not affected by the test chosen, the starting point of the deletions, or the rate of the deletions - as long as it is greater than five and less than ten words (pp. 219-220). In other words it is commonly assumed that deletion pattern does not influence cloze test results. Klein-Braley reviewed previous studies and found that authors rarely questioned the equivalence of cloze tests. She feels that evidence for this is the widespread failure to provide reliability figures, suggesting that the absence of such data indicates the authors believe it is not required. Indeed, another area where further study is needed is the reliability of cloze test scores.

In the majority of cases the reliability coefficients reported are in the region of .9 and thus indicate that the reliability of the cloze tests involved was extremely satisfactory. Some researchers, however; report coefficients which are considerably lower than .9 (cf., eg., Enkvist and Kohonen, 1978; Alderson, 1978a .) (p. 221)
Klein-Braley notes that correlation coefficients are usually KR-20. The KR-20 is a measure of internal consistency calculated from a single administration of a test. She states that in order to establish the consistency of results derived from cloze tests we need evidence collected from two (or more) cloze tests administered to the same examinees." (p. 222). In "an extensive survey of the literature" Klein-Braley found that only two studies mentioned such data; Anderson and Porter. Anderson (1976) gives satisfactory correlations between cloze tests which are "in general higher than .9" (Klein-Braley, p. 222), though his study dealt with very young children and may not apply to adults. Porter's (1978) results were unsatisfactory with all correlations falling below .65 (Klein-Braley, p. 222).

As Klein-Braley has noted, few writers have compared the results from more than a single administration of a cloze test. It appears that more work is needed particularly of a cooperative nature involving several researchers using the same test over multiple administrations.

A review of research on cloze procedure has shown that further research is needed to establish the most appropriate method of scoring cloze tests and that such a study should compare the results of different cloze tests administered to the same examinees.
CHAPTER THREE

Introduction

This section contains a description of the way in which the study was conducted. It describes the development of the tests used as well as the population which served as subjects.

A pilot study was conducted to check on possible problems which might arise during the administration and scoring of cloze tests. A small class from Vancouver Community College was selected on the basis of convenience. Vancouver Community College is a college which provides ESL training to immigrants mainly, though French Canadians also study there. The students varied in age from 20 to 48 and they came from a variety of cultural and language backgrounds. The first sentence of the passage selected was left intact and every fifth word was then deleted until the cloze tests were fifty items in length.

Students were given as much time as they required to finish. Checking the tests after administration revealed that all items had been completed, indicating that the test was probably not a speeded test. A speeded test is one which provides only enough time to be completed by the most proficient student. Most of the students completed their papers in about thirty minutes.
The tests were then scored by matching each response with the original deletion (exact scoring) or judging items as being appropriate in their surrounding context (acceptable scoring). A graph of the results is seen in Figure I.

**FIGURE I Comparison of Exact and Acceptable Scores (Pilot)**

The first observation was that acceptable scoring results in higher scores, a fact noted by earlier writers. It was found that though the results produced by each scoring method were similar when the entire set of scores was
considered, it was possible that variation might occur in the case of particular individuals. The second observation on the graph, for example, shows that the student was ranked as lowest by the acceptable method and second to lowest by the exact method.

Having reviewed the results of the pilot study a new test was designed. Different passages were used in order to test for passage effect. The length of the test used was doubled and two passages were included on each test resulting in a test containing two complete 50 item tests based on different passages; a technique employed by Porter (1978, p. 335). Thus each test would be composed of two 50 item cloze tests; one passage on each side of a single sheet of paper. One passage would be common to each test paper and would thus serve as a sort of benchmark. The first sentence was left intact and deletions began in the next sentence. Passage B was from Kisslinger and Rost (1974, p. 38) beginning "Most industrialized nations . . ." and ending "There was a lot of it and it was easy to get." Passage A was taken from Alexander and Wilson (1974, pp 24-25) beginning "I have always quite irrationally prided myself . . ." and ending ". . . and drive about everywhere in it." Passage C was taken from Dixson (1971, pp. 49-50) beginning "Some words have interesting origins." and ending "They were in a panic." The
passage not common to every test would randomly alternate between two different passages, a design which has several positive features. In doubling the test by including two passages, it is possible to administer two cloze tests to each examinee. Since three passages are used (one common to all tests the other two alternating randomly), it is possible to examine the effect of using different passages, at least to some extent.

Each test contained one passage - call it "B" which was common to all tests and an additional passage selected from two alternatives - "A" and "C." In order to minimize both the effect of the order of presentation (if a passage is first or last) and the proficiency of the examinee, four test forms were created:

**FIGURE II  Test Forms**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>page 1</td>
<td></td>
<td>page 1</td>
<td>page 1</td>
<td>page 1</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>page 2</td>
<td>page 2</td>
<td>page 2</td>
<td>page 2</td>
<td>page 2</td>
</tr>
</tbody>
</table>
These forms were then assigned the numbers 1, 2, 3, or 4. A computer program was then used to generate a random permutation of the numbers from 1 - 4. The printout of the result was then consulted and the test papers were stacked accordingly. The result was a stack of test papers randomized according to the computer printout.

Prior to administration of any tests the teachers of particular classes were contacted. In some cases TOEFL scores for students were available and these were collected in order to be used for later comparisons.

Columbia College and Vancouver Community College were the two sources of examinees. They are colleges in the city of Vancouver, Canada. Both provide ESL programs to immigrants who need to gain proficiency in order to successfully function in an English speaking society. The examinee population may have differed in several ways from the generally more uniform population which seem to have been tested in much previous research.

During administration of tests examinees were ensured that their performance would not be counted against them. They were also informed that, as this was a study, they were not required to participate. This announcement was made in keeping with the rules governing research utilizing human subjects that are required of research at the University of British Columbia and elsewhere. In announcing that results are not to be
taken into account in final grades, it possible that performance of examinees will differ from their performance on a "real" test. Of course, since students are in the environment of a new country and are dependent on the administration of the school, it is likely that participation (or at least giving the impression of participation) was viewed as prudent.

Papers were handed out with instructions to fill in the information on the top of the test (name, age, etc.). Talking was discouraged as much as possible. In cases where it appeared that two people were exchanging information, their papers were checked when they were handed in to see if there was evidence of copying. Sufficient time was given to ensure that the test was not a "speeded" test. It was found that one hour was ample time, even for the slowest people.

The tests were scored by three people; myself and two untrained volunteers who were not teachers. Several hours were spent pondering the criteria to use in scoring responses according to the acceptable method. In the end the instructions chosen were simply "if it looks appropriate to you, mark it right." This definition of "acceptable" may seem tremendously vague, but in this statement's vagueness lies its strength. It was assumed to be easily understood by native speakers and flexible enough to allow acceptance of not
only synonyms of the original deletion, but any response which would be
acceptable by a native speaker. In designing this study an attempt was made
to behave like "real" teachers do. Teachers don't (in most cases) spend
tremendous amounts of time developing scoring procedures for their tests. In
fact, one of the main positive features of cloze mentioned by several writers is
its simplicity.

The data collected were entered into the mainframe computer at the
University of British Columbia and calculations were made using BMDP\(^2\).
The data was also keyed into an Apple Macintosh Plus and analysed using
the program Statview 512+\(^3\). Pearson product moment correlation
coefficients were calculated in order to compare exact scoring to acceptable
scoring.

\(^2\) BMDP STATISTICAL SOFTWARE, INC.
1964 WESTWOOD BLVD. SUITE 202

\(^3\) BRAIN POWER, INC.
24009 VENTURA BLVD.
SUITE 250
CALABASAS, CA 91302
CHAPTER FOUR

Introduction

This chapter will describe the results obtained from the administration of cloze tests in this study.

Ninety-two papers were collected. Since each examination paper contained 2 cloze tests, data for 184 cloze tests were obtained. However, since only one passage was common to all papers, the largest number of scores collected for one passage was 92. The other passages were randomly distributed; combined together making a total of 92. Forty-five papers contained passage "A" and 47 papers contained passage "C." These numbers are very close to the ideal of 46 each, which would make each passage 50% of the total.

Population

The population of examinees at first appeared to be extremely diverse. Several first languages were represented: English, Chinese, Punjabi, Malay, Persian, Japanese, Swahili, Kileyyu, Urdu, German, Philipino, Spanish, Arabic, Greek and Russian. However, when the actual numbers were checked, it became clear that the population was made up of primarily speakers of Chinese.

Three examinees did not indicate their first language. Of the remaining 89,
60 indicated Chinese was their first language. The next most frequently listed first language was English (7) followed by Malay and Persian (both 4). Arabic and Japanese were both represented by two examinees. All other languages were represented by only a single examinee. The large number of Chinese speakers in the sample may be related to Canadian immigration at the time these data were collected. Figure III shows the representations of the various languages.

**FIGURE III** Pie Chart of First Language

The examinees were spread through 7 different classes taught by 4 different instructors. Ages ranged from 15 to 25. There were 17 examinees
aged 19, 16 aged 20 and 15 aged 18. See Figure IV.

**FIGURE IV Box Plot of Age**

If the total scores are considered for all the papers some general features become apparent. First, the scores for the acceptably-scored cloze tests are, overall, higher than those for the tests scored according to the exact response method. This difference is clearly seen in the box plot below. The five horizontal lines on the box indicate the 10th, 25th, 50th, 75th and 90th percentiles respectively. The small circles represent individual scores lying outside the 10th and 90th percentiles.
If the two different scoring methods do produce different results, then not only will one set of scores be larger than the other, there will be a difference in the distribution of scores. In other words, there would be not just a quantitative difference, but a qualitative difference as well. If such a difference does exist, the two sets of scores would not be expected to correlate highly. If, on the other hand, the two different scoring methods produce substantively similar results, then a high correlation would be expected. If all test scores are considered, the correlation between Exact and Acceptable scores is extremely
high (.92). This coefficient is much higher than those usually reported in the literature but it is based on tests which used two cloze passages; creating a test double the length of a "normal" 50 item cloze test. This increased length may be responsible for the very high correlation coefficient; though of course, a longer test cannot increase the strength of a correlation unless it exists in the first place. It is also the case that the total scores are derived from various combinations of three passages. A similarly high correlation was reported by Irvine et al. who give a correlation of .94 between exact and contextually acceptable scores. Their cloze tests were only 50 items in length. If the same correlation is calculated for all the first pages of the tests used in this study (one cloze passage only) the correlation is reduced but remains high. In the case of all page ones the correlation is .89, for all page twos the correlation is .82. These somewhat lower coefficients are more similar to those reported previously. The smaller magnitude of the correlation is likely to be the result of a smaller pool of results, since each page is a single 50 item cloze test. Remembering that the scores for all the cloze tests contain the results from various passages, it will be of some interest to check passage B, which was the only passage common to all 92 cloze tests. When exact and acceptable scores are compared the correlation is .87; still an acceptably high result.
There are still two passages to consider. Passage A and Passage C represented a total number of 45 and 47 examinees respectively. In passage A the correlation between the exact and acceptable scores was .83. Exact and acceptable scores correlated at .88 in passage C. It appears that the two methods of scores produce quite similar results. The high correlation between exact and acceptable scores indicates that the increase in scores obtained on the acceptably-scored papers does not result in any change in results. In other words, with acceptable scoring the distribution of scores is higher but places individuals almost identically with respect to exact scores. This similarity between scores is seen if one consults a graph of the two sets of scores. If the only difference between the two scores is that one is higher than the other, it will be possible to increase all the exact scores by some specific amount which will make them identical to the acceptable scores. In order to test this, a graph was constructed by subtracting the difference in means between the two sets of scores and adding this difference to each exact score. Altering the scores in this manner allowed the same scale to be used for each set of scores. The results were then plotted jointly with the scores obtained with the acceptable method of scoring. See Figure VI next page.
FIGURE VI  Acceptable Scores and Exact Scores with Mean

Difference Between Scores Added

Line Chart for columns: $X_1 - X_2$

- Mean Diff. + Total Exact
- TOTAL AC

Observations

61
The similarity between the two sets of scores is striking. Since it is possible that there is some variation in results in individual passages, it is necessary to compare exact and acceptable-scoring results between the three passages. Comparing passage B with passage C, the exact scores correlate at .74 (N=47) and the acceptable scores at .66 (N=47). Passage A exact scores correlate with passage B exact scores at .53 (N=45) and the acceptable scores correlate at .66 (N=45). It is not possible to determine the correlation between passages C and A, since students completing passage C did not complete passage A.

Comparison With TOEFL

Fortunately several students (38) had been previously tested with the Test Of English as a Foreign Language (TOEFL). Unfortunately the TOEFL scores had been collected, in some cases, "a few months before." Due to the time spanning between the administration of the TOEFL and the cloze tests developed in this study, it is possible that any differences in results are due to changes in the performance of the examinees.

When the TOEFL and cloze tests are compared the correlation between Total Exact Score and TOEFL is .65; between Total Acceptable Score it is .67. The correlation between scores is also seen in the scatter plot of the scores
in Figure VII. The individual passage scores correlate with TOEFL as follows.

Passage B exact scores: .65 N=38  
Passage B acceptable scores: .65 N=38

Passage A exact scores: .55 N=18  
Passage A acceptable scores: .68 N=20

Passage C exact scores: .71 N=20  
Passage C acceptable scores: .67 N=20

FIGURE VII Scattergrams for TOEFL vs Acceptable and Exact Scores

Scattergram for columns: $X_1Y_2$  R-squared: .449

Scattergram for columns: $X_1Y_1$  R-squared: .422
CHAPTER FIVE

Introduction

This chapter contains a brief review of this study, its findings and the implications of these findings.

Several writers have suggested that cloze procedure is "an accurate and reliable tool" (Jonz, 1976, p. 255) for the measurement of overall language proficiency. Despite an enormous amount of material written in favor of cloze procedure, others suggest that more research is needed before cloze procedure is used as a measure of proficiency in classrooms (Klein-Braley in Oller 1983; Porter 1978). Klein-Braley suggests that what is needed in particular are studies of cloze procedure involving different cloze tests administered to the same students. It is also clear in reading research on cloze procedure that scoring cloze tests is a matter of some controversy. While Oller and Conrad (1971) favor the counting of "any contextually acceptable word" (p. 191) as correct, others suggest that it is sometimes difficult to determine what responses are "acceptable" (Kryzanowski, 1976). Miller and Coleman stated that exact response scoring should be used, since allowing acceptable responses did not result in substantially different scores (1967). It is clear, after a review of the research concerning cloze procedure, that more
Research is needed, particularly investigations of different methods of scoring cloze tests utilizing different cloze tests with the same subjects. Therefore, the focus of this study was a comparison of two methods of scoring cloze tests, since it is crucial that the most reliable method be used in order to assess students' performance. Since the most appropriate scoring methods for general use are exact and acceptable scoring, these scoring methods were selected for study.

In order to obtain as much information as possible, two cloze tests were included in each test. One passage was common to each test, the other was chosen from two other passages, thus giving two possible sets of passage. A deletion rate of every fifth word was used since it was in keeping with previous studies and it also permitted one cloze passage to fit on a single page allowing a single sheet printed on both sides to contain two independent 50 item cloze tests. A single page kept cost to a minimum and was much more convenient to handle than two separate sheets. These two combinations of passages were then presented in alternate formats (different order of presentation) creating four separate test forms. These different forms were then randomly administered to 92 students and then collected and scored according to two methods. The exact replacement method was used first.
Each paper was scored by matching the student's response to that on a key indicating the original word deleted. The result was then tallied for each page, since each page was an individual 50 item cloze test. The same test was then re-scored by determining if the student's response seemed appropriate to the person marking the test. Three different people scored the tests, each scorer completed approximately 30% of the papers. This procedure gave two scores for each cloze passage; the exact and acceptable scores. The two sets of scores were then compared in order to determine whether or not the two method of scoring produced similar results. This was done by correlating the two sets of scores. It was found that total exact scores (for all passages) correlated at .92 (N=92) with acceptable scores. This correlation is high and indicates that there is little difference between the two methods. The correlations between the acceptable and exact method for each passage were: passage A .83 (N=45), passage B .87 (N=92) and passage C .88 (N=47). These results indicate that, for the same passage, the exact scores and acceptable scores correlate at over .8 indicating that one set of scores can be substituted for the other with little loss of information. The situation is not the same when different passages are compared. All page ones correlated with all page twos for the exact method at .44 (N=92) and for the
acceptable method at .62 (N=92). These correlations are lower, particularly for the exact method, indicating that scores obtained from different passages may not give the same distribution of scores. A comparison of all page ones with all page twos involves a comparison of scores obtained from two different passages for the same subject. If Passage A is compared to passage B, the exact scores correlate at .53 (N=45) and acceptable scores correlate at .66 (N=45). If passage C scores are compared to those from passage B, exact scores correlate at .74 (N=47) and acceptable scores correlate at .66 (N=47). Since different students completed passages A and C, it is not possible to compare the scores from these passages. It is interesting that in the comparison of passages A and B, the highest correlation is obtained with acceptable scores, while in the comparison of passages C and B, the highest correlation is obtained with exact scores.

A comparison with TOEFL scores for the various passages indicated varying degrees of correlation.

- Passage B exact scores: .65 N=38
- Passage B acceptable scores: .65 N=38
- Passage A exact scores: .55 N=18
- Passage A acceptable scores: .68 N=20
- Passage C exact scores: .71 N=20
- Passage C acceptable scores: .67 N=20
The correlation between cloze and TOEFL scores ranged from .55 to .71. These correlations may not seem particularly high, but given the change in proficiency which is likely to have occurred between the administration of the two tests (TOEFL and cloze) and also the small number of scores available for comparison, it seems to indicate that there is some relationship between the two sets of scores. If it is assumed that TOEFL is an acceptable measure of language proficiency, then a high correlation between cloze tests and TOEFL scores is desirable. Unfortunately, due to the elapse of time between the collection of the TOEFL and cloze scores it is not possible to make any conclusion concerning the correlation between the scores.

Limitations of this study

In order to get the best results possible, the weaknesses of previous studies were examined in order to avoid them wherever possible. The main weakness of previous studies appears to be that they derive their results from a very specific population; a problem often exacerbated by having a small number of subjects in the study. Porter (1978), for example, based his study on results obtained from 39 students (p. 335). Similarly, though Alderson (1983) administered his tests to 360 subjects, only 30 subjects completed a single test form (p. 207). Unfortunately, despite good intentions, this study
cannot be said to have a large number of subjects. Constraints of time and availability of classes meant that only 92 tests could be collected. This number is somewhat larger than several other studies, particularly if one considers the number of subjects per test form. Though the present study is significantly different from other studies who have relied on university students for subjects, there is still much room for further work utilizing a much larger population of subjects.

Another weakness of this study is that, despite an attempt to get a population consisting of a variety of language backgrounds, the examinees were mainly speakers of Chinese. It is possible that the results may have been somewhat different with a more diverse population.

A similar point can be made concerning the use of different passages. Just as it is important to study a diverse population, it is important to determine if the passage selected will have any effect on cloze scores. Unfortunately, to really study a significant number of passages requires that several tests be administered to an extremely large population. Using different passages created another problem. If several different tests are used the number of any specific test drops. For example, if 100 people are available to be tested and four different tests are used, there are only 25 people writing any particular test. Klein-Braley (in Oller, 1983) also encountered this problem:
... only two cloze tests... have the usually recommended 50 items. In performing this study it was necessary to choose between administering one full-length 50 item cloze test and two shorter tests, since all tests were administered together with other test procedures and the time available for the cloze tests was limited. Again, however, this is not unusual (cf. eg., Pike, 1973, n=25; Wijnstra and Van Wageningen, 1974, n=45; Enkvist and Kohonen, 1978, n=43. (p. 225)

Having a small number of any particular test reduces the strength of any statistic calculated for that test. Given the rather limited availability of classes it seemed best to use fewer passages, thus obtaining more data for the passages chosen, rather than selecting a large number of passages and collecting limited data from each. It is quite possible that results may have differed had other passages been selected. On the other hand, if cloze tests are extremely sensitive to changes in passage, such sensitivity would likely be revealed in this study. In fact, the scores obtained by the same examinee on two separate passages were quite different. That is, the correlation between scores on the two halves of each test was not high, indicating that different passages did influence performance on cloze scores.

An important area of investigation which was not adequately dealt with in this study was the actual process of acceptable scoring, that is, what do scorers do when they make decisions about the "correctness" of a replacement word? It is important to note, in a study of cloze in particular, that
there is a difference between what is supposed to happen and what happens.

Near the completion of this study it was found that, though the relationship between exact and acceptable scores suggests that there is high inter-rater reliability, the scorers sometimes marked responses in unexpected ways. In the following sentence: "By burning wood people were able to __________ their homes" the two responses "warm" and "build" were given by two different students. The response "warm" seems more correct than "build", since burning wood is more directly related to warmth than construction. Both were considered acceptable by two different scorers. In the sentence: "Wood is a __________ fuel . . .," the response "endless" was marked correct by one scorer, even though it is clearly inappropriate, since a word which begins with a vowel is never preceded by "a." Clearly more work is needed in discovering just what scorers do when they score cloze tests.

In this study the inter-rater reliability of cloze scores was not calculated since the required data were not collected during this study. This information would have been interesting in view of the discovery that scoring was producing some unexpected results.

**Implications of Results**

The focus of this study was to determine the most reliable method of
scoring cloze tests. The data collected indicated that exact scores correlate very highly with acceptable scores. This is very interesting, particularly when the acceptable method used in this study would, at least intuitively, maximize chances of increased unreliability. Not only were three different raters used, the definition of acceptability was kept purposely vague in order to emulate a "real" test as much as possible. If it is indeed difficult to determine what an "acceptable" response is, as Krzyzanowski (1976) states, one would expect that acceptable scores would not correlate well with exact scores. This was not the case. Further, utilizing three scorers rather than one would tend to increase any discrepancy in scores, since it is possible different scorers would have different opinions of what constituted "acceptable" responses. In spite of these factors, the correlation between acceptable and exact scores was quite high, indicating that each scoring method produces similar results.

It would seem, based on the data presented in this paper, that the exact method of scoring could (or should) be used instead of acceptable scoring, since there is little difference in the information provided by the two sets of scores. The exact system of scoring is frequently recommended on the basis of its simplicity.

Although slightly higher reliability has been obtained at other times using other procedures, such as a synonym count, the increased time and subjectivity necessary for such systems do not
warrant their use. (Jongsma quoted by Rupley, 1973 p. 496)

The acceptable scoring done in this study, in contrast to the extreme complexity and strain suggested by Jongsma was accomplished with only minor expenditure of time. Two of the scorers were not teachers, and despite their inexperience in scoring tests of any type, found scoring by the acceptable method to be a relatively simple task. Acceptable scoring in fact, is easier than exact scoring in that it does not require a comparison between the student's response and the originally deleted word. The responses are scored on the basis of one's proficiency as a native speaker (or at least as a very proficient speaker). To state that a proficient speaker of English is unable to score according to the acceptable method seems to call into question the kind of proficiency that is being required of the examinee. If the proficiency being demanded of a student of English is even beyond the capability of someone regarded as fluent, then such tests should not be administered. On the basis of the experience of scoring the cloze tests in this study, acceptable scoring is not only easily employed, it is correlates quite highly with the exact response method. When arguments of fairness are considered, the acceptable method of scoring seems more suitable in calculating scores.

Once it is accepted that it is quite possible to have exact word replacement with little or no understanding of the text, that the exact word need not necessarily be the "best" word in terms of
precision, expressiveness, etc., or the only suitable word, and that the requirement of exact word replacement may introduce bias into the test in favor of those whose socio-economic position gives them greater access to western culture in general, or native speakers of English in particular, the arguments for acceptable word scoring become very strong indeed. (Johnson, 1981, p. 196)

Acceptable scoring can be recommended on the basis of a much simpler reason than that supplied by Johnson. Acceptable scores are higher than exact response scores. If exact scores are calculated, even the most proficient students will be unable to get nearly perfect scores. If proficiency in a language is viewed as something that native speakers possess, then a test which measures this proficiency should allow native speakers to achieve perfect or nearly-perfect scores; cloze tests scored according to the acceptable method seem to be closer to this view.
Conclusion

In conclusion, this study has provided evidence that the exact method of scoring and the acceptable method of scoring cloze tests correlate sufficiently highly to allow the use of either method. In cases where the scorer is not a proficient speaker it may be best to use the exact method of scoring. In cases where the scorer is a proficient speaker, it may be most appropriate to use the acceptable method, since it results in higher scores, allowing native speakers to achieve nearly perfect scores. It was found during the course of this study that determining the "acceptability" of responses was not a difficult task. A positive point in favor of the exact method should be mentioned: it may prevent students and teachers alike from taking scores too literally, since such scores seem intuitively unfair in many cases. Some caution in interpreting test scores as measures of real proficiency is desirable.

The results of this study also suggest that some caution should be taken in the selection of passages for use in cloze tests. It appears that although cloze procedure per se may provide a reasonable method of testing, individual cloze tests may not. It is probably advisable to select a set of passages and test them in a pilot study in order to determine their suitability for inclusion in a final test.
A more fundamental consideration in using cloze procedure is its appropriateness in the situation in which it is being used. If the comments by Ebel and Carver (1961, p. 641) are accepted, cloze procedure may be most appropriately used as a measure of proficiency, perhaps as a screening test, than as a test of the amount of learning of classroom skills. The "meaningfulness" of cloze procedure must be determined by the tester. As was pointed out in a discussion of the validity of cloze procedure, this is not a simple task. Each test administrator must determine the appropriateness of cloze tests within their own classrooms. In situations where a measure of overall proficiency is needed, cloze procedure may be a useful technique for developing tests. Based on the findings of this paper, the choice of acceptable or exact scoring can be made to suit the circumstances of the testing situation; they are, at least in terms of reliability, interchangeable.

Areas for Further Research.

Further research is needed in order to determine how effective cloze procedure is as a measure of overall language proficiency. More work must be done to determine how various factors, such as student proficiency, deletion rate, passage, etc. influence cloze test results. More work is also needed in order to determine what it is that scorers do when they determine whether or not a student response is "acceptable".
References:


Aitken, Lewis R. 1985. Psychological Testing and Assessment. Allyn and Bacon, Inc. 7 Wells Avenue, Newton, Massachusetts 02159.


Longman Group Ltd. London


78
Carroll, J. B., A.S. Carton, and C.P. Wilds. 1959. An Investigation of Cloze Items in the Measurement of Achievement in Foreign Languages. Cambridge, Massachusetts: Laboratory for Research in Instruction, Graduate School of Education, Harvard University. ERIC ED 021


