A Study of the Cloze Procedure with Native and Non-native Speakers of English

John Charles Alderson

Ph. D.

University of Edinburgh

1978

I hereby declare that the thesis has been composed by myself and is entirely my own work.

John Charles Alderson

Abstract

This study examined various aspects of the methodology of the cloze procedure to determine their effect on the nature of cloze tests. It was hypothesised that changes in the frequency of word deletion, in the difficulty of the original text and in the procedure used to judge acceptable restorations of the deleted word would produce significantly different cloze tests and would result in varying correlations with measures of English proficiency.

Three texts were selected and each was subjected to the deletion of every sixth, eighth, tenth and twelfth word, to give twelve cloze tests. Five procedures were developed to score the responses to these tests for the degree of similarity they showed to the deleted words.

The tests were administered to 360 adolescent native speakers of English and 360 adult non-native speakers of English who were pursuing further studies in Britain.

It was found that significant differences existed among cloze tests when deletion frequency was changed, but that some scoring procedures reduced this effect. The change in deletion frequency had no effect on the measurement of text difficulty, but significant interactions were observed among the three experimental variables. Different cloze tests gave unpredictably different measures of English proficiency. A study of identical deletions showed that no increase in the predictability of deleted words was gained by increasing context from five words to eleven words.

Since the quantity of context had no effect on predictability, it was suggested that cloze is essentially sentence-bound. The nature

of the correlations of cloze with measures of English proficiency and the results of factor analyses suggested that cloze is a better test of syntax and lexis than of higher-order reading abilities. Implications for future use of the cloze procedure are presented and suggestions made for further research.

Contents

Volume I: The Study

Volume II: Bibliography, statistical tables, graphs, and appendices

2.1 Relating to Chapter 4 - Algerian pilot study

2.3 Relating to Chapter 6: Results 1) : Native Speakers

2.4 Relating to Chapter 7: Results 2) :

Non-native Speakers

C.H A P T E R 0

Introduction

0.1 Uses of the cloze procedure

It is a quarter of a century since Wilson Taylor published an article in Journalism Quarterly in 1953 on a procedure for measuring the readability of text by randomly removing words from that text, replacing them with a standard-length blank, and requesting subjects to attempt to restore the deleted word. This procedure, which he called the "cloze procedure", subsequently aroused great interest, and there has been a remarkable flourishing of proposed uses of the procedure. In the United States, the number of doctoral dissertations written either on the cloze procedure or its use as a tool, has risen from two in 1967 to twelve in 1973, and fourteen in both 1974 and 1975.

The uses to which the cloze procedure has been put are varied. It was initially used to measure the readability of texts, comparable with traditional readability formulae such as the Flesch or the Dale-Chall formulae, although it was soon claimed to be superior to them in many respects. It has been used to measure the readability of elementary algebra textbooks, mathematical English, electronics texbooks, social studies materials, business letters, poems, headlines and telegraphic prose. It was quickly assumed that the procedure, in addition to measuring texts, could also measure the characteristics of the reader, and by 1957 it was being claimed that cloze measured the reader's degree of comprehension of text. As a measure of such comprehension it has been used with normal readers at almost all educational levels, from primary through to postgraduate university students. It

has also been employed to compare the comprehension abilities of deaf and hearing children (e.g., Odom et al., 1967) and the educable mentally retarded (e.g., Semmel et al., 1967).

As a supposed measure of linguistic predictability the cloze procedure has been applied to transcripts of the speech of alcoholics, of schizophrenics and of aphasics. It has been used to examine the language of psychiatric interviews, and to establish the high predictability of working-class "restricted code" speech as compared with the low predictability of middle-class "elaborated code" speech (Poole, 1972). It has even been claimed that a cloze test on a transcript of recorded speech can show that marihuana may interfere with the retrieval of information from the brain's immediate memory store (Weil and Zinberg, 1969).

Apart from written text, the cloze procedure has also been used with tapes of spoken discourse as a measure, among other things, of listening comprehension, and one investigator (Lynch, 1972) used the technique on visual images to measure audience response to cinema films.

The last decade in particular has seen a growing use of the cloze procedure with non-native speakers of English to measure, not only their reading comprehension abilities, but also their general linguistic proficiency in English as a Foreign Language. The technique is now widespread throughout the English teaching world, particularly in the Third World, and is used by classroom teachers to construct tests of their students' linguistic abilities which they assume to be valid measures of such abilities.

Nevertheless, some researchers and teachers have wondered whether the cloze procedure really is a valid measure of all that is claimed for it, and whether, perhaps, the technique needs closer investigation, especially in view of its current popularity, to attempt to establish its nature and value. Indeed, it was largely the increasing popularity of the cloze technique that occasioned the present study. The views of some researchers as to the usefulness of the technique will be presented in Chapters 1 and 2, after a fairly comprehensive survey of the literature, and the main body of this work will present a series of investigations into the nature of the cloze technique, both with native and non-native speakers of English.

0.2 What is the cloze procedure?

By cloze technique is meant the use of random or selective deletions of words from continuous text to measure reading or listening comprehension. This is not the same as the "Sentence Completion Technique", where half of a sentence (beginning or end) is removed and the subject expected to replace what he thinks has been removed. The literature on cloze sometimes includes reference to the sentence completion technique, whose applications are many and varied. It has been used in psychiatry to identify neurotics and psychotics, in psychology to differentiate aggressive types from non-aggressive types, to measure parents' attitudes to teachers, as a means of studying interpersonal development, as an exercise in verbal operant conditioning and, in linguistics, as an elicitation technique.

The essential difference seems to be that the sentence completion technique is intended to elicit data in a controlled way,

but does not overly concern itself with comprehension of the given text, or with its nature. The sentences used are written by the experimenter for their ability to suggest ideas to the subject. Completion of the sentences is said not to involve closure, but creation. The researcher is not interested in the relation between the responses and what was deleted, nor in the degree of uncertainty about the deleted word (its "entropy"). Cloze technique is said to involve closure based on the predictability of the text and the expectancies of the reader, and, therefore, to depend heavily on comprehension of the preceding text. It is said to be a measure of that text and its contextual interrelations, and the researcher is interested in the relationship between the response and what was deleted, as well as the entropy of the blank. The sentence completion technique is interested in independent points of information, not contextually interrelated ones.

However, the random or selective deletion of words can be undertaken in various ways, and for this reason the term "cloze procedure" can be used to refer to at least three hierarchically different procedures. At the most general level, the term simply refers to the systematic removal of words from a text, to be replaced by subjects and scored in some non-specified manner. This definition covers any use whatever of the procedure.

At a less general level, it is possible to interpret "systematic" in different ways. One interpretation is "random" or "pseudo-random", and in this procedure words are removed from text either by reference to a table of random numbers, or by an every-nth

word procedure, where deletion begins with, say, the first word, and then every eighth word thereafter, regardless of its nature, is removed to construct a cloze test.

Another interpretation of "systematic" is that the words to be deleted are selected according to some rational principle, based on the nature of the words themselves. Thus it is possible to delete only those words thought hard to replace, or those words which are felt to be highly redundant. More usual is the deletion of words from certain linguistic categories. For example, only nouns are deleted from text, or only function words, or verbs and adjectives. This procedure is usually referred to as a "rational" cloze procedure, as contrasted with a random or pseudo-random procedure. (Other terms exist in the literature, for example, Rankin (1957) refers to rational cloze as "lexical deletion" and random cloze as "structural deletion".)

Thus, at the second level of generality, the term cloze procedure would refer to either rational or random deletion of words from text.

An even more specific use of the term "cloze procedure" is sometimes encountered in the literature, and this use refers to the random deletion of every fifth word from text, thus excluding any other deletion frequency.

The three possible definitions of the term "cloze procedure" are thus: at the most general level, systematic deletion; at a less general level, either random or rational systematic deletion; and at the least general level, pseudo-random systematic deletion of every fifth word.

When writers talk of cloze scores, then, it may be to any of these three definitions that they refer. Unless otherwise stated, this work will refer to the cloze procedure as defined at the most general level, that is, including rational and random deletion selection procedures.

There is another dimension to be regarded when talking of cloze scores, and this is the dimension of scoring for correctness. Traditionally, there are two types of cloze scores. The first one accepts as correct only the actual word deleted from the passage. Any deviation from this word (other than minor spelling errors) - such as morphological, syntactic or semantic change - is regarded as incorrect. This score is known as the "exact word" (or "verbatim") cloze score, and is the most commonly used. However, another score is also used by researchers, in which any word is accepted as correct which is either a synonym of the deleted word, or which "fits into the context". (Clearly, criteria for acceptability may vary.) These procedures are known as the"synonym" or "any acceptable word" scoring methods respectively.

Other scores are occasionally reported. One is to accept as correct any word which comes from the same linguistic form class (noun, preposition, etc.) as the word deleted - this is known as a "form class score". Another score is the "communality of response score", sometimes also referred to as the "clozentropy method" if calculated in a specific mathematical way, where the subjects' responses are checked for their agreement, not with the word deleted, but with the responses provided by some criterion group of subjects.

It can thus be seen that when a writer refers to the "cloze procedure", he may have in mind the random deletion of every fifth word from text, and the exact replacement of the word deleted (admittedly the most usual meaning of the term, especially in the United States). Or he may have in mind the deletion of every second verb and adverb from text, where replacements are given credit for the amount of agreement they have to responses provided by superior adult readers. It seems clear that the reader of reports of the cloze procedure should ensure that he knows which procedure is being referred to, and it is also clear that caution must be applied when trying to treat all these different techniques as one "cloze procedure", since it is possible that what holds for one cloze procedure does not necessarily hold for another.

0.3 What the cloze procedure is said to test

When Taylor introduced the cloze procedure in 1953, he presented a series of semi-theoretical justifications for its validity. The word "cloze" itself was created from the Gestalt concept of closure, that is, a tendency for humans to form a complete whole by filling in gaps in a structure. Just as there is a tendency to see a not-quite-complete circle as a complete circle by closing the gap, making the image conform to a familiar shape, so the subject in a cloze test is said to "close" or "cloze" by linguistically completing an incomplete structure.

In order to be able to "cloze" the gap, Taylor postulates, the subject must know the meanings and the forms of most or all the words involved, and the meanings of the combinations of both in a given sentence structure. In other words, one must "understand" the muti-

lated sentence as a whole, and then complete its pattern. If reader and writer share the same or similar language habits, the reader should be able to make accurate restorations of deleted words and thus the cloze procedure becomes a measure of the similarity between the patterns that the decoder is anticipating and those that the encoder has used. As Wilson and Carroll (1954) put it:

> "If the encoder producing a message and the decoder receiving it happen to have highly similar semantic and grammatical habit systems, the decoder ought to be able to predict or anticipate what the encoder will produce at each moment with considerable accuracy. In other words, if both members of the communication act share common associations and common constructive tendencies, they should be able to anticipate each other's verbalisations."

Thus the cloze procedure is justified not only as a measure of closure, but also as an objective measure of the language correspondence between reader and writer.

Since cloze scores are apparently affected by at least the variables of reader, writer and text, it follows that the procedure can also be used to measure characteristics of messages. Because subjects are able to restore mutilated text, it is claimed that cloze is a measure of the redundancy of that text. Linguistic redundancy is seen as the extent to which the information presented by morphemes is recoverable from other parts of the utterance containing those morphemes. Thus some morphemes can be absent and the information they carry can be recovered from the remaining text. For example, "man coming" means the same as the sentence "A man is coming this way now".

The latter contains redundant elements: it indicates the singular number of the subject three times - indefinite article "a", singular noun "man" and singular verb concord "is"; the present time reference twice - present tense "is" and adverbial "now"; and the direction of action twice - "coming" and the phrase "this way". Such repetitions presumably make for ease of restoration if one of the redundant elements is deleted. The ease with which words from a cloze test can be restored can thus be taken to be a measure of its redundancy.

Redundancy is enhanced in text by transitional probabilities in language, which have the effect of increasing predictability. Some transitions from one word to the next are more probable than others: thus "A HAPPY NEW" is more likely to be followed by "YEAR" than "CHRISTMAS" or "CAR". It can thus be claimed that to the extent that cloze measures the predictability of text, it is also a measure of transitional probabilities. To the extent that transitional probabilities reflect language habits and are higher in familiar sequences, a measurement of transitional probabilities also reflects the comprehensibility of text, or the ability of the reader to understand text.

Such are the traditional rationales offered in explanation of what it is that a cloze test tests.

Taylor (1953) has this to say about the cloze procedure:

> "A cloze score appears to be a measure of the aggregate influences of all factors which interact to affect the degree of correspondence between the language patterns of transmitter and receiver. As such, its potential usefulness is by no means confined either to readability or the reading abilities of individuals."

The rationale of the random cloze procedure is simply that if enough words are randomly deleted, the blanks will represent proportionately all kinds of words occurring in that text, and will thus represent an adequate sample of the linguistic difficulties contained in the text.

The following three chapters present a survey of the research into and with the cloze procedure to date, with native speakers of English (Chapter 1) and non-native speakers, mainly of English (Chapter 2). Chapter 3 presents a more detailed survey of some aspects of the cloze procedure which might affect its nature. Chapter 4 describes an experimental pilot study of some of these aspects, while Chapters 5 to 8 contain the description of the main study. Finally, in Chapters 9 and 10 the results are discussed, and their implications are considered.

CHAPTER 1

The Use of the Cloze Procedure with Native Speakers
(A Survey of the Literature)

1.1 Pre-Taylor

Although the idea of a gap-filling exercise was known long before Taylor, its use was quite different from that proposed by him. Ebbinghaus (1897) was interested in such a technique (which he called the "Kombinationsmethode") for the measurement of intelligence, and his studies were further developed by people like Brown (1910), Ballard (1920) and Hamilton (1929). None of these people were interested in the technique as a measure of either the readability of text or of the reading comprehension of subjects, but rather as a general measure of verbal intelligence. Their methodology differed from that of the cloze technique in that they selected words to be deleted from sentences and paragraphs according to their notions of the substitutability of those deletions; the substitutions provided by the respondent were then examined for the light they could throw on that person's intelligence. For this reason, the Kombinationsmethode could more appropriately be termed a "sentence completion technique" (see Chapter 0).

1.2 Taylor

Taylor (1953) was interested in the value of a technique of mechanical deletion of words from text as a measure of the readability of that text. The assumption he made was that words removed from difficult passages are harder to replace than words deleted from

easier texts. He hypothesised that cloze scores would rank the three passages he had chosen in the same order as the two readability formulae - Flesch and Dale-Chall; that the difference among the three passages would be significant; and that the relationship between the cloze scores would remain the same despite the use of different word deletion systems, different presentation orders, and different scoring procedures.

Trying several deletion systems in his first experiment - removing every fifth word, removing every tenth word, and removing ten per cent of the words randomly - he found that they all discriminated significantly among the passages, and in the same order as that predicted by readability formulae. Although both the every-fifth deletion system (which removed 35 words) and the random-10% procedure (which removed 16 words) discriminated better among the six subjects than did the every-10^{th}-word procedure, Taylor claimed that random and every-n^{th} deletion systems would give equivalent results if more than 16 words were deleted from the text. Moreover, since his concern was not subject discrimination but passage discrimination, he concluded that any deletion system would give the same results.

He also compared a scoring procedure which allowed only replacements of the exact word deleted as correct with a procedure which also gave partial credit (half a point) for "good enough" synonyms of the deleted words and found that although slightly higher mean scores were achieved for the passages with the second procedure, the degree of differentiation among passages was identical with either procedure.

Taylor's first experiment also established that the order in

which the three passages were presented to subjects had virtually no effect on scores.

To validate the results of the first experiment, Taylor selected five more passages which were to be examined along with the original three. The new passages were taken from texts which Taylor considered to be extremely difficult, but which readability formulae would classify as being relatively easy (from Gertrude Stein and James Joyce), and, conversely, from texts which the formulae would classify as difficult, but which he considered to be easy. He thus hypothesised that the cloze procedure would be able to predict the "true" readability of the passages (as judged by Taylor) more accurately than the readability formulae. This time he deleted every 7^{th} word from each of the eight passages. His hypothesis was confirmed that cloze rankings of passages correspond closer to the intuitive rankings of difficulty than do the formulae rankings. In fact, no significant correlation was found between cloze and either of the formulae. He also found, by examining consecutive groups of five items, that the same ranking of passages would have been achieved if only half as many words had been deleted from each passage. (Moreover, he discovered that high and low scoring subjects were separated from each other within the first five items of the cloze tests.) Interestingly, he also found that a pilot study group of six subjects predicted the relative cloze difficulty of the eight passages in exactly the same order as did a subsequent independent sample of 18 subjects.

Taylor concludes his paper by saying that, although the cloze procedure seems to be a valid and reliable measure of readabi-

lity, and possibly even of reading ability, more research is required to make the procedure as efficient as possible, and to validate it extensively. The effect of manipulable factors should be examined in properly controlled experiments. One example of these manipulable factors is the deletion system employed, and Taylor suggests further experimentation to determine whether an every-5th deletion rate is more or less efficient than an every-10th- or every-15th-word deletion, and also to determine how many deletions would be required for dependable results.

Taylor himself followed up some of his suggestions in a subsequent study (Taylor, 1957), which not only investigated the methodological considerations of different deletion systems, but also examined the validity of cloze indices of readability by determining the degree to which the cloze scores of individual subjects correspond to measures of specific knowledge, comprehension and general aptitude. The notion was tested that a person's cloze score on a passage would be a measure of his mental ability, how much he knew of the article's content before studying it, and how much he knew after reading. The criterion of mental ability was the Armed Forces Qualification Test (AFQT) which had subtests on, inter alia, word knowledge and arithmetical reasoning. The criterion for knowledge of subject matter before and after studying an article was a homemade multiple-choice test given before and after study. Scores on these three criteria were compared with individuals' scores on one of three types of cloze test given before reading the article on which the test was based (the "before" cloze), and again after reading the unmutilated text, seven days later (the "after"

cloze). One version of the cloze was the (subsequently usual) every-5th-word deletion procedure, and the other two versions deleted "easy" and "hard" words respectively - where "easy" and "hard" were determined both by linguistic analysis, and from previous cloze data. Scoring of responses was by the exact word procedure only.

Taylor found that the cloze procedure correlated significantly with his measures of comprehension, with coefficients ranging from .51 to .92. However, not only did the "before" cloze tests correlate substantially with the "before" comprehension test (from .58 to .92), and the "after" cloze with the "after" comprehension (from .64 to .80), but also the "before" cloze tests predicted the "after" cloze tests (from .80 to .88).

The cloze tests also correlated positively with the AFQT (from .46 to .74) at approximately the same level as the comprehension tests. Further, Taylor found that the differences between "before" and "after" test scores, for both cloze and multiple-choice tests, were always significant, so that the cloze could be said to be measuring not only comprehension, but also learning.

Although correlation coefficients and score gains were significant for all three deletion systems, the "easy" word deletion generally differed from the "any" and "hard" deletion systems, such that it correlated considerably lower with the three criterion tests, and gain scores were also lower.

Differences between "any" and "hard" deletion patterns appeared only in the correlation coefficients: in general, the "any" deletion procedure correlated as high as or higher than the "hard"

system with the criterion measures, with the sole exception of the pre-test of knowledge, which correlated higher with the "hard" word deletion.

Taylor concluded both that the cloze is a reliable and valid measure of comprehension, specific subject knowledge and general aptitude or ability to understand, and that, of the three deletion systems, the "any" form, which is the simplest to construct, yielded more stable, reliable and discriminating results than did the "easy" and "hard" forms.

Thus the two most important initial studies of the cloze concluded that it was a reliable and valid measure, both of text readability, and of subjects' reading abilities, and drew certain methodological conclusions.

1.3 Cloze as a Measure of Text Readability

Taylor's use of the cloze procedure to measure the readability of text for native speakers of English has been followed up many times in the last twenty years. One of the more important early studies of cloze in this respect was by Bormuth (1963). Bormuth hypothesised that there was a correlation between the cloze test difficulty rankings of a set of passages, and their readability levels as measured by multiple-choice tests. He produced nine passages, three in each of the subject areas of literature, social studies and science, to the Dale-Chall readability levels of 4.5, 5.5 and 6.5; subjected the passages to a deletion of every fifth word; and administered the tests to groups of children from school grades 4, 5 and 6. He found that each of the grade level groups agreed on the relative difficulty

of the passages (rho: .99, .98 and .98); in other words, that the cloze was a reliable measure of readability for all three levels, and, further, that the difficulty levels as established by the cloze tests agreed with the difficulty level established by the multiple-choice tests (rho: .92, which was as high as the reliability of the tests would allow). He concluded, therefore, that cloze is a valid and highly reliable predictor of the comprehension difficulty of the passages.

Mosberg, Potter and Cornell (1968) carried out a similar study, with subjects at the 5^{th}- and 8^{th}-grade levels. They selected passages from the SRA laboratory materials, at difficulty levels either two years above, two years below, or at the subjects' grade level. Every fifth word was deleted from these passages, and the cloze procedure was compared with multiple-choice tests on the same texts. They found that whereas the cloze tests showed no differences among easier passages, the multiple-choice tests were sensitive to the differences, showing a linear decrease in performance over difficulty levels. They suggest that the cloze lacks "sensitivity at the lower end of the difficulty dimension".

Nevertheless, their findings went relatively unnoticed and research proliferated with the cloze procedure as a criterion measure of readability, for a great variety of subjects and texts. With subjects at grades 9 and 15, Froelich (1970) investigated the readability of electronics textbooks, and found that the cloze procedure identified the readability levels of the texts "in a manner more consistent with the abilities of college students" (as measured by an

achievement test) than did the Flesch Reading Ease Formula. In fact, the Flesch formula did not predict the observed difference in cloze scores on the texts, and thus it was concluded that cloze was the superior measure of readability.

Hater and Kane (1972) examined the reading difficulty of passages written in mathematical English, taken from instructional material intended for grades 7 to 12, by comparing cloze tests on the passages with multiple-choice comprehension tests, and discovered that the two methods were in close agreement on the relative difficulties of the texts. It was, therefore, concluded that cloze tests can be used as predictors of reading difficulty for mathematical English content.

Froese (1971) compared cloze with the Dale-Chall readability formula as measures of the readability of 6^{th}-grade science textbooks, and obtained a correlation between the two of -.29. He therefore concluded that the Dale-Chall formula is not a valid measure of such materials when the cloze is used as a criterion, and, by implication, that the cloze is a valid measure.

In Britain, Moyle (1970) looked at the cloze as a readability measure for young children (age 6 to 10), and concluded that it was a better measure than the Fry readability graph, but that, on the whole, the difficulty ranking established by cloze agreed with that of the publishers.

Moyle points out, however, that there are a number of aspects of the cloze on which further information is required, and these include the equation of a cloze score to so-called independent

and frustration reading levels, the effect of deletions of particular parts of speech, and the appropriate number of deletions for each age group. As will be seen later, the first two questions were already in the process of being investigated, but the third question had barely been considered, and indeed remains uninvestigated.

Jefferson (1969) examined the effect of three deletion types on cloze as a measure of readability. He used the random every-5th-word procedure and compared it with deleting every fifth noun versus every fifth adjective, and deleting every fifth structure word. He found a significant interaction between readability and deletion type, and suggested that readability as measured by the cloze is related to the type of deletion such that passage difficulty rankings are not the same when using n^{th} deletion scores, and when using scores where the deletions are certain categorized language variables.

Two recent studies of readability have tried to find the most appropriate deletion rate for measuring readability. McNinch, Kazelski and Cox (1974) asked whether a single deletion pattern is appropriate for all subject matter areas, and attempted to investigate the relationship between deletion patterns, passages and reading abilities. They used four texts from the subject areas of English, science, social science and mathematics, and subjected them to three deletion patterns - the deletion of every 5^{th}, 7^{th} and 9^{th} words. Although their results revealed a significant effect of deletion pattern, they found no consistent best pattern for a subject matter. However, they conclude that science materials need a low frequency deletion, and that social science and English materials should have a deletion rate

of every 7th word. Unfortunately, their evidence does not back up this finding, which will be discussed at greater length in Chapter 3, but the consequence for studies of readability is that a cloze test will not necessarily give an appropriate measure of the readability of a particular text, since estimates of the readability vary according to the deletion pattern used.

A similar conclusion was reached by Miller and French (1974), who examined deletion rates 5, 7 and 10 as measures of the readability of science and social science material. They used passages at four different grade levels, as determined by readability formulae, and appear to have found that whilst the subjects were able to read the social science materials equally well, regardless of deletion pattern, the science material was only read "adequately" at the 7th-word deletion frequency. However, they suggest that the best deletion pattern for measuring readability might be the every-10th-word rate.

It can be seen that the trend in the use of cloze as a measure of readability has been from an initial validation of the technique against existing readability formulae and multiple-choice comprehension questions, to a point where the cloze, now regarded as a valid measure of readability, has been used as a criterion with which to compare other readability measures, usually to find them lacking. In its heyday, the cloze was used to measure the readability of almost any written matter, from telegrams to captions, from headlines to academic textbooks. However, there has recently been a gradual realisation that perhaps the cloze technique is not necessarily a reliable measure of readability, since the use of different deletion patterns

may well affect the estimate of text difficulty.

1.4 Cloze used to measure linguistic variables

Closely linked with these readability studies has been the use of the cloze technique to investigate the linguistic variables that affect the readability of text. Bormuth has been concerned with the investigation of such linguistic variables, and their incorporation into new readability formulae for some time. In 1964, he reported on a study (Bormuth, 1964a) using the cloze technique as a criterion of readability. Cloze tests on nine texts from the areas of literature, science and social science were used to establish the readability of the passages. He then calculated the Mean Word Depth score for each passage. (Word depth is based on Yngve's hypotheis (Yngve, 1960) that sentences whose immediate constituent structure is right-branching will be easier to understand than left-branching structures. He developed a system of counting such left-branching structures to give an index of sentence difficulty.) He found that although there was a correlation between mean word depth and the Dale-Chall readability rating of 1.00 when holding subject matter constant, the mean word depth correlated .78 with the cloze criterion of readability when the Dale-Chall effect was reduced to zero. He claimed, therefore, that mean word depth can predict differences in comprehension difficulty (assumed to exist because of the cloze scores) among passages that differ in subject matter, but whose sentence length and proportion of hard words are almost identical, and concluded that sentence length is not an adequate parameter of readability.

In a series of articles (Bormuth 1964b, 1966, 1968a),

Bormuth reported on an extensive investigation into readability and linguistic variables, using cloze to determine the comprehension difficulties of twenty passages, and of each word, independent clause and sentence within each passage. Subjecting the passages to different linguistic analyses he produced new linguistic variables which can be used to develop new readability formulae which will be better predictors of passage difficulty. Some of these variables are word depth, the number of letters in a text and the ratio of different parts of speech, especially, unaccountably, the ratio of pronouns to conjunctions. He claimed that improved formulae are now possible because, whereas previous formulae had been validated against multiple-choice comprehension scores on the passages under investigation, it is now possible to use the cloze technique as a criterion. This technique, moreover, according to Bormuth, gives an estimate of the readability, not only of text as a whole, but even of individual words in that text.

Coleman (1971) also reports on an investigation into the relationship between various linguistic variables and the cloze scores of texts. Interestingly, he comes to a conclusion similar to Bormuth's, in that he finds the highest correlations with the cloze ranks of text difficulty are achieved by counts of letters, syllables and morphemes in the texts, such that "passages become harder to understand as they contain more letters, syllables or morphemes." Nevertheless, he also observes a considerable increase in the predictive power of linguistic variables "as one progresses from relatively gross syntactic units (for example, sentences) to more refined ones (for example, kernel sentences)".

Many studies have been made of the relationship between cloze and specific linguistic variables, in an attempt to uncover not only what constitutes difficulty in text, but also in order to investigate the cloze itself, and to investigate the linguistic variables. Taylor (1953) recognised that cloze items could be classified as easy and hard, and that this division corresponded with a rough division into function and content words respectively. Aborn, Rubinstein and Sterling (1959), in a study that used sentence completion rather than cloze, systematically deleted words from sentences and examined the effect of form class on the difficulty of restoration. They concluded that the predictability of a word is inversely related to the size of its class. This, of course, is essentially the same finding as Taylor's, since there are few members of the function word classes, and large numbers of words in the content word classes. In particular they found that adjectives and adverbs had a low predictability, and that when responses to deleted adjectives and adverbs were made, they were rarely in the same form class.

A similar finding was made by Fillenbaum, Jones and Rapoport (1963) to the extent that they found adverbs to be hardest to replace by a word from the same form class. Nevertheless, in general they found relatively few differences in the predictability of the form class for the different form classes, even when comparing function words with content words. Of course, content words were harder to replace exactly, and so they hypothesised that form class predictability might be more dependent on the relatively close grammatical environment, whereas verbatim predictability might depend more

on both close and remote semantic features of the discourse. Thus, different deletion rates might be expected to have different effects on form class and verbatim cloze scores. However, they also point out that the form class of some members of a particular form class is much more predictable than that of other members. To some extent this depends on the form classes (and therefore, presumably, on the grammatical environment) of the words preceding and following the deletion. Thus in the environment "adjective - blank - verb" a noun might be easily predicted.

Luke (1964), in a study of form class and cloze procedure, investigated 10-percent deletion of four types: nouns only, verbs only, adjectives only, and combinations (five each of nouns, verbs and adjectives). The highest mean scores - i.e., the easiest class to replace - were achieved by verbs, followed by combinations, nouns, and adjectives. (Here the scoring was by the exact word only.)

Louthan (1965) carried out a somewhat different study. He subjected text to seven different types of deletion: 10% of nouns, verbs, modifiers (adjectives and adverbs), prepositions and conjunctions, noun determiners, pronouns, and every tenth word regardless of class. He then tested comprehension by means of questions rather than by requiring subjects to restore the deleted word. He found no difference in comprehension when deleting every 10^{th} word, nouns only, verbs only or adjectives only, but it was more difficult to answer questions on text with these deletions than on undeleted text. Curiously, however, he also found that when only prepositions and conjunctions were deleted, or when only pronouns or determiners were removed, superior

comprehension scores were achieved over the control groups reading undeleted text. Although the researcher does not conclude that removing function words from text makes it easier to understand, he does claim that his results show that content word deletion requires different skills from those required when function words have been deleted.

A similar study, deleting selected word types and testing comprehension by means of questions, was carried out by Bradley (1969), who found that the deletion of nouns had a significant effect for all his groups, and that the deletion of adjectives had an effect for one group, but that deleting verbs and function words had no effect. Unfortunately, the results were vitiated by the finding that when no text was present, subjects could answer the multiple-choice test at a level greater than chance.

Like other researchers, Rentel (1969) discovered that adjectives were hard to predict, but he also found that as the words to be deleted increased in length, the difficulty of predicting them increased significantly across all form class categories. He concludes that for all words over seven letters long, length is the most important determiner of difficulty, not form class. This, of course, is related to measures of readability used by the most common formulae, and to the well-known fact that long words tend to be more difficult than short ones.

Tannenbaum, Williams and Clark (1969) found that if they told people doing a cloze test what the form class of each deletion was, not only did their ability to restore with a word of the correct form class

increase, but also their verbatim score for function words increased. However, their verbatim score for content words did not change. Reasonably, they suggest that this is probably because of the small number of items per class for function words compared with the large number of words in each content word class. However, it should be noted that if the information leading to the correct form class identification of function words is already available in the context of the deletion, then the additional information supplied by the tester should be entirely redundant, and therefore have no effect on either form class or verbatim scores. That this is not so indicates that the context does not contain information sufficient unambiguously to assign the form class even to a function word.

At this point it is useful to recall Coleman's paper (1971), already referred to, in which he examined the effect of various word classes, but pointed out that

> "The most important general conclusion seems to be that traditional definitions of word classes are too imprecise and too gross to yield profound insights into verbal behaviour. Some definitions of a word class correlate positively with comprehensibility; others correlate negatively. The major word classes contain subclasses that correlate positively and other subclasses that correlate negatively."

Not all investigators of the cloze and linguistic variables have confined themselves to looking at form classes. Coleman and Bormuth, as has been seen, looked at the effects of various variables on comprehensibility. Ruddell (1965, a and b) investigated the effect of writing passages using the linguistic patterns that occurred

frequently in the oral language of the target population, compared with texts written using low frequency patterns. (An example of the latter might be: "The leader gave the men short breaks because they needed rest", whereas the former might be: "A spaceman could fix the small hole.") Vocabulary and sentence length were held constant. He discovered that cloze scores on high frequency pattern material were significantly higher than scores on low frequency pattern material, and this he attributes to the greater structural redundancy of the easier passages. Thus readability is a function of the redundancy of the syntactical elements of the materials or, put another way, "of the similarity of patterns of language structure in the reading material to oral patterns of language structure used by children."

Coleman and Blumenfeld (1963) compared the cloze scores of nominalizations and their grammatical transformations using active verbs, and discovered that the word patterns of the passages written in active verbs were considerably more predictable. Fagan (1971) found that cloze scores were generally lower for grade three children when embedding and deletion transformations were present than when conjoining or position shift transformations were present. Peltz (1974) found that cloze tests showed passages repatterned according to the linguistic patterns used by subjects in their own writing to be easier than the original passages, and many of these linguistic patterns involved simple, expansion and conjoining transformations.

In a different type of study, Darnell (1963) examined the effect of rearranging sentences in text to create seven varying degrees of disorder. He found that disorder can adversely affect com-

prehension as measured by cloze, and that the amount of loss of clarity becomes greater as the degree of disorganisation increases.

Finally in this survey of the use of the cloze procedure to measure readability and investigate linguistic variables, Ohnmacht and Fleming (1974), using sentences rather than texts, investigated three different verb types: 1) a transitive verb, 2) a complement verb with direct object ("decide", "believe"), and 3) a complement verb with no direct object, but a clause ("He believed in doing his best"). Although their analysis revealed a significant effect of verb type, the hypothesis that type 1) would be easier than type 2) would be easier than type 3) was not confirmed. Interestingly, the main finding was that bilateral constraint - i.e., words of context both sides of the deletion - had a great effect on ease of restoration.

1.5 Cloze as a measure of reading comprehension

Taylor's 1957 paper showed that cloze correlated well with comprehension tests on the same material (from .51 to .92) and, in particular, his "any-word" deletion (what has come to be known as the random or every-n^{th} cloze procedure) correlated better with other measures of comprehension than did his "rational", "easy" or "hard" word deletions.

In the same year Rankin (1957) reported an extensive investigation into the cloze, in which he discovered that a pre-cloze test, rational deletion (a pre-cloze test is a cloze test given to subjects without their reading the unmutilated text first) correlated .59 (corrected .86) with a comprehension test on the same passages,

whereas a post-cloze test (after reading the intact passage) correlated .56 (.78 corrected) with the same test.

Jenkinson (1957) compared cloze test scores with scores on specially written multiple-choice comprehension tests based on the same texts and obtained a correlation of .82.

Bormuth (1962) wrote a 31-item comprehension test, carefully controlled for vocabulary, and obtained correlations from .73 to .84 with individual cloze tests and a grouped cloze score that correlated .93 with the multiple-choice test.

Bormuth (1969) constructed a battery of tests, on the same passages as his cloze tests, to measure vocabulary, facts, sequences, relationships, main ideas, inferences and the author's purpose. He achieved a range of correlations, from .35 to .89, with the majority of coefficients in the region of .80.

Most studies come to the same conclusions about the validity of cloze as a measure of the comprehension of a text. However, one study (Mosberg et al., 1968) introduces a note of caution. Their correlations between cloze and multiple-choice tests ranged from .43 to .65 for a fifth-grade group of subjects, but only .19 to .37 for an eighth-grade group. Correlations averaged out over three cloze passages were .54 for the former group, and .34 for the latter. They conclude that "a large component of comprehension as measured by multiple-choice tests is not accounted for by the cloze procedure", and, further, that

> "one should proceed cautiously when using cloze test scores as measures of comprehension (or predictors of multiple-choice test

score magnitudes) until a more detailed analysis of what the cloze procedure is measuring has been established."

Many studies have sought to determine what it is that the cloze procedure measures by correlating close test scores with standardised tests of various sorts. The commonest criterion test is some test of comprehension or verbal ability. In general, correlations of the cloze with such tests are lower than correlations with tests based on the same passage as the cloze test, but nevertheless they are quite high and positive. Rankin (1957) obtained correlations of .68 and .60 with the Vocabulary and Paragraph Comprehension sections of the Diagnostic Reading Test, Survey Section, although with Story Comprehension the correlation was as low as .29. However, rational clozes, and cloze tests given after reading the undeleted text once, increased the correlation for Story Comprehension and reduced it for Vocabulary and Paragraph Comprehension.

Jenkinson (1957) correlated an every-n^{th} cloze test with the Cooperative Reading Test and obtained coefficients of .78 with Vocabulary and .73 with Level of Comprehension. Using the same criterion test, Fletcher (1959) got lower correlations with the Vocabulary section (.63) and Level of Comprehension (.55), whilst his cloze test correlated .57 with the Speed of Comprehension measure and .59 with the Rate of Comprehension section of the Dvorak-Van Wagenen Diagnostic Examination of Silent Reading Abilities.

Ruddell (1963), using five cloze tests, obtained a range of correlations (from .61 to .74) with the Paragraph Meaning subtest of the Stanford Achievement Test.

Gallant (1965) used primary grade children as subjects, unlike most other studies, and achieved correlations ranging from .65 to .81 for her cloze tests and the Metropolitan Reading Test.

Greene (1964), like Rankin, correlated cloze tests with the Total Comprehension score of the Diagnostic Reading Test, Survey Section, and obtained coefficients of .51 for a random cloze procedure, and .67 for a modified rational procedure. He also constructed his own tests to measure "words" and "relationships between words" and found significant correlations of between .49 and .59 with his cloze tests. Unlike many observers, he concluded from these results that a considerable amount of variance was not accounted for by cloze scores, and that the cloze procedure is more complex than previously assumed.

Hafner (1964) looked at the relationship of various measures to the cloze and came to the conclusion that reasoning is important in doing cloze tasks, since he obtained correlations of .73 with the Otis Quick Scoring Mental Ability Test, and .56 with the Information subtest of the Wechsler-Bellevue Intelligence Scale. The finding that ability to do cloze is quite closely related to measures of intelligence merely confirmed previous findings by a variety of researchers. In fact, Taylor (1957) had found a correlation of .85 with the Word Knowledge subtest of the AFQT, and .70 with the Arithmetical Reasoning subtest of the same battery. Rankin (1957) found that a structural cloze test (every n^{th} word deleted) resulted in a fairly high correlation with measures of intelligence (.68 and .60), whereas lexical cloze (nouns and verbs only deleted) was less closely related to intelligence (.32 and .21) and more to what he calls "pure" comprehension.

Jenkinson (1957) obtained a correlation of .69 between a random cloze test and intelligence quotients, whilst Fletcher (1959) found a correlation of .72 with the Linguistic subtest of the American Council on Education Psychological Examination, again an intelligence test. Deutsch et al (1964) found that cloze scores were significantly related to IQ, as did Schneyer (1965), who echoed Rankin's (1957) suggestion that although rational cloze was significantly related to IQ, it was less related than a random cloze procedure (in his case, deletion of every 10^{th} word). Greene (1964) obtained correlations of .52 and .61 between two types of cloze test (random and modified rational) and Thorndike's test of verbal reasoning. Finally, Froese (1971) found that cloze scores correlated from .55 to .85 with the Canadian Lorge - Thorndike IQ test.

In addition to studies of cloze as a measure of intelligence, the procedure has also been investigated as a measure of various other abilities, not so closely related to specific reading comprehension. Bormuth and MacDonald (1965) claimed that completing the cloze, that is, "exactly matching an author's words, requires an acute sensitivity to literary style, choice of words, sentence patterning, attitudes towards his subject matter, and aesthetic devices", and in fact found correlations of between .45 and .66 with a test of ability to detect literary style. Ratekin (1971) examined the adequacy of the cloze to measure comprehension of different logical patterns (induction and deduction), whilst Byrne, Feldhusen and Kane (1971) reported a relationship between cloze and three measures of divergent thinking abilities.

Simple correlations have not been the only statistical tool

used to investigate what the cloze procedure measures. Several investigators have availed themselves of the relatively sophisticated procedure known as factor analysis to try to isolate the factors underlying the relationships between cloze and other variables. One of the first such analyses was carried out by Weinfeld (1959, reported in Carroll et al, 1959), who investigated 28 cognitive tests, one of which was a normal, random cloze test. He found that the cloze was related to a number of cognitive factors, and reported the association as: Ideational Fluency .40, Word Fluency .50, Fluency of Expression .54, Theme Writing .59, Verbal Factor .70, Reasoning Factor .76.

Weaver and Kingston (1963) used eight different cloze tests in a battery of 18 cognitive tests, which included tests like the Davis Reading Test, the MLAT Number Learning Test, tests of sensitivity to grammatical structure, rote memory, sound-symbol association and the like. Their factor analysis revealed three main factors, which they labelled Verbal Comprehension, Redundancy Utilization and Rote Memory. Unexpectedly, the cloze had only a low loading on the verbal comprehension factor and on rote memory, whilst having a high loading on redundancy utilization (essentially a cloze factor). They showed that cloze tests are even more related to each other than to the other two main factors revealed. Unlike many investigators since, who have been disturbed by this lack of correspondence between cloze and verbal comprehension, Weaver and Kingston were not discouraged by their findings, since "unless the cloze procedure can be shown to elicit variance from some source <u>other</u> than those of more commonly used reading and language objective tests, its use is likely to remain more of an interesting curiosity than a valuable research and measurement tool". (They over-

looked the convenience of the procedure, which has proved so attractive to subsequent researchers and test makers.)

However, Bormuth (1969) severely criticized the Weaver and Kingston study, for using college students as subjects, and for using comprehension tests based on different texts from the cloze tests. However valid his criticism, his own factor analysis, based on nine cloze tests and seven tests of reading comprehension on the same texts, designed to measure vocabulary, facts, sequences, relationships, main ideas, inferences and the author's purpose, resulted in only one factor with an eigenvalue greater than unity, which accounted for 77% of the variance. This factor, called reading comprehension ability, was very closely related to the cloze tests, so he concluded that cloze did indeed measure reading comprehension.

The other major factor analytic study was made by Ohnmacht, Weaver and Kohler (1970), with a replication with older subjects in 1972. Testing the hypothesis that cloze measures closure, they used a battery of tests of speed and flexibility of closure, as well as tests of associational fluency and verbal comprehension. They also constructed four different types of cloze tests: structural (every 5^{th} word), lexical (every 5^{th} noun, verb and adjective), abstract (every 5^{th} abstract noun) and concrete (every 5^{th} concrete noun). The intercorrelation of these cloze tests was somewhat low, between .4 and .5. Three factors emerged: a general cloze factor, followed by a perceptual factor, and a verbal ability factor. Cloze had only low loadings on the perceptual (closure) factor, whilst closure had virtually no loading on the cloze factor. The lexical and abstract clozes were the only cloze

tests to have moderate loadings on the verbal ability factor, thus supporting a theory that distinguishes between random and rational (or structural and lexical) cloze procedures. Indeed, the authors recommend that rational deletion patterns should be investigated more closely and tied in with linguistic and psycholinguistic theory. The replication study (Ohnmacht and Fleming, 1972) came up with similar general factors, but found fewer significant correlations among tests, and concluded that whilst cloze may be factorially complex, the relationship of verbal and perceptual abilities with cloze factors may be moderated by the grade level of subjects. Nevertheless, both studies cast some doubt on the theory which relates cloze to closure, and on the relationship between cloze and verbal comprehension, and also support other findings (see Chapter 3) that different cloze deletion types may be related to particular linguistic and psycholinguistic abilities. "Cloze", it is suggested, is not a unitary concept, and the ability to do "cloze" is probably complex.

1.6 Cloze and multiple-choice test scores

Despite the doubts expressed by some researchers as to whether cloze really measures verbal comprehension, other investigators have been so encouraged by the correlations often achieved between cloze tests and more traditional comprehension tests that they have tried to equate cloze scores with multiple-choice scores on the same text. Bormuth (1967) was the first person to present a table of correspondence between the two types of scores, where he shows that a cloze score of 38% is equivalent to a multiple-choice test score of 75%, and that a cloze score of 50% is equivalent to a multiple-choice

score of 90%. The importance of these levels is that they are said to indicate the limits of the "study" level of comprehension. In other words, a person scoring between 75% and 90% on a multiple-choice test on a passage is supposed to be capable of reading that text for study. A score of above 90% is said to indicate the student's capacity to read the text *on his own*, whereas a score below 75% represents a level of comprehension likely to lead to frustration. Although Bormuth included caveats on the interpretation of his cloze scores, largely (and sensibly) related to the cloze methodology, his remarks have been largely ignored in practice.

Rankin and Culhane (1969) replicated Bormuth's study, using 5th-grade students, and corroborated his finding, equating a multiple-choice score of 75% to a cloze score of approximately 41% and a multiple-choice test score of 90% to a cloze score of 61%.

1.7 Reading gain

The term "reading gain", as used by Rankin (1957), refers to the amount of information acquired by a reader from a text. This acquired information is not the same as the information available to a subject after reading, since whereas the latter information depends on the amount of information the reader had before reading the text, "reading gain" is independent of this. It is usually measured by giving the same test before and after reading a text, and subtracting the difference to arrive at a gain score.

There is some evidence that cloze can measure such gain. Taylor (1957) found significant differences in scores from cloze tests given before and after reading text. Rankin (1957) also found

significant increases in post-cloze test scores over pre-cloze scores. Moreover, he found that unlike the pre- and post-cloze test scores, the gain score was not related to intelligence.

However, doubt has been cast on the ability of cloze to measure such reading gain by a study carried out by Coleman and Miller (1968). Using matched subjects, they gave one group a cloze test on a text, and had the other group read the text before taking the cloze test. A correlation of .93 was found between the two tests, and no significant gain was achieved. The authors conclude that the cloze test given before reading a passage is an inadequate measure of how much a subject knows about a passage before reading it, and that it is measuring essentially the same information as the cloze test after reading. The conclusion seems not unreasonable in view of the fact that the subject has to read the text in order to complete the cloze, which is the reason for the claim that it is a measure of comprehension. Therefore the subject is presumably capable of gaining information whilst doing the cloze.

Bormuth and Macdonald, in an earlier study (1965), had also discovered that there was no difference between cloze tests given before studying an author, and tests given after studying the author and the books from which the cloze tests were constructed. However, they suggested that the reason for the difference between their study and Rankin's (1957) might be that they had used a normal random cloze, whereas Rankin had used a lexical, rational deletion system.

The validity of a gain score was questioned by Taylor in Greene (1967), where he reports a study of a cloze test given to sub-

jects to complete, and then given again two weeks later, with no intervening reading or information on how they had performed on the test. He discovered a significant increase in scores and concludes that his subjects must have learned during the initial cloze test.

1.8 Cloze as a teaching device

A great deal of time and effort has also gone into experiments designed to prove that the cloze technique is an effective teaching technique (for improving reading comprehension skills). Smith (1969) used cloze exercises for improving reading comprehension in a junior college reading program, and claimed that it was particularly effective for demonstrating the process of comprehension - "how words combine with words into wider units of meaning" -, for demonstrating the part grammatical knowledge plays in comprehension, for pointing out to students their own deficiencies in the comprehension process, and for promoting discussion about particular reading selections and the process of "reasoning which is reading comprehension". He does not appear to have proved this experimentally, however. Nor has scientific research borne him out. Phillips (1973) showed that cloze exercises did not improve the reading ability or content achievement of junior college students in an Introduction to Business course, compared with students who had no such training. Ellington (1972) found no significant difference between cloze exercises, conventional exercises and no exercises to improve reading comprehension, as measured by standard tests. Faubion (1971) found no proof that training in cloze resulted in a growth in comprehension skills (although he recognises that there may have been certain variables which

influenced the effectiveness of the cloze training). Kennedy (1971) found an increase in comprehension after cloze training as compared with oral reading practice, when measured by standard tests, but not when measured by cloze tests. Rynders (1971) failed to find any significant difference between cloze exercises and traditional comprehension exercises as a method of increasing reading comprehension, but the author admits the possibility that the standard test used may have failed to detect differences that did exist. Friedman (1964) tried with foreign students but, again, found no difference between cloze exercises and close reading, although the subjects reported that they were more aware of the structures when using cloze exercises. Jongsma (1971) concludes his survey of the use of cloze in teaching by stating that the research evidence does not suggest that cloze, as it is presently understood, is an effective teaching technique, and proposes that research be carried out into the effect of varying the deletion system.

1.9 Summary

The general consensus of studies into and with the cloze procedure for the last twenty years has been that it is a reliable and valid measure of readability and reading comprehension, for native speakers of English. Investigation of the readability of text first justified the use of the cloze procedure by showing that it compared well with standard readability formulae. They then showed it to be superior to such formulae in many cases, and the cloze procedure was frequently used as a criterion measure of readability against which to gauge new and developing measures of readability. As an extension of

its role in readability studies, the cloze has been used to investigate the linguistic difficulties of text, in particular the predictability of particular word classes, and the difficulties caused by various types of sentence complexity.

As a measure of the comprehension of text, cloze has been shown to correlate well with other types of test on the same text, and also with standardised tests of reading comprehension. Moreover, it has also been seen to correlate well with IQ tests, and other tests of cognitive abilities. It has also been suggested that the cloze technique can be used for measuring the information a reader gains during the reading process. The procedure has been used in exercises to increase reading comprehension abilities, though with limited success.

1.10 Doubts about the cloze

However, not all studies have shown the cloze to be capable of all that is claimed for it. Some researchers have always expressed doubts about the validity of the use of cloze as a measure of comprehension, others have doubted whether the procedure can be used accurately and consistently to measure the reading difficulty of text, and yet others have wondered whether what has been claimed for one type of cloze is necessarily true for all types.

The original claim by Taylor that the cloze requires the Gestalt operation of closure was challenged by Weaver (1965). He claimed that, far from being an essentially perceptual or matching operation (which is what he claims the reading process is), the cloze is a cognitive process, similar to a coding operation. To fill in a cloze test, the subject must make a search of the distribution of

probable elements suitable in that environment. This search procedure is not a logical exhaustive process, but rather a heuristic procedure, Tuinman and Blanton (1971) support this rejection of Taylor's original rationale for the cloze with experimental evidence showing that the distribution of cloze responses is related to success probability.

Whatever the nature of the psychological processes involved in cloze, it is sometimes claimed that cloze is not a measure of comprehension of text. Weaver and Kingston (1963) seemed to show that cloze was unrelated to comprehension. MacGinitie (1971) claims that subjects are often capable of restoring words successfully in a cloze test with only a recognition of "familiar patterns of expression", and no true understanding of the text. Cloze, he claims, is less a measure of comprehension and more a measure of redundancy.

However, Bowers and Nacke (1971) show the inadequacy of information theory to account for linguistic facts (for example, creativity, or infinite recursiveness) and so claim that a procedure like cloze, which, they say, is based on information theory, is also "of dubious worth in testing the linguistic behaviour of a language user". Since linguistic redundancy does not operate in the way that information theory predicts, a cloze procedure which randomly deletes every n^{th} word is incapable of measuring the redundancy of text. Linguistic redundancy is determined on syntactic and semantic grounds, and to tap such redundancy a cloze procedure would need to make selective deletions based upon a linguistic analysis of the redundancy of that text.

Brown, in Greene (1967), suggests that there are various

hierarchically ordered levels of comprehension, which he calls "sampling, matching, accepting, interpreting, understanding and believing". He further suggests that the exact word method of scoring cloze only taps up to the fourth level of comprehension, that scoring for synonyms will enable the fifth level to be measured, but that there are levels of comprehension which are not tapped by the cloze.

Rankin (1974), in an article reviewing the use of the cloze procedure over the previous twenty years, echoes a growing feeling of concern over the nature and validity of the cloze procedure. He suggests that perhaps the cloze is a better measure of readability than of reading comprehension. He criticises much work on the procedure for using the random procedure rather than the rational deletion of words from text. (In this, he echoes a remark in MacGinitie (1971) to the effect that rational deletions are more likely to measure comprehension than linguistic redundancy.) This concentration on random cloze, he claims, has strengthened the influence of general verbal abilities and intelligence upon the cloze measurement of reading comprehension. After listing many limitations of the cloze procedure, including the facts that many cloze items are not determined by context and so do not discriminate in testing terms, that many items are not "reading related" but "reflect background information or general language ability", that many items depend on short-range linguistic constraint, that perhaps the deletion of every 5^{th} word is not the most suitable deletion pattern for every use, and that perhaps rational deletion of words from text is no longer a cloze procedure, Rankin

comments on the mushrooming use of cloze in recent years and adds a word of caution:

> "Performance on a cloze test . . . is influenced by the reading ability of the reader and the difficulty of the materials, (and) . . . the type and number of items deleted. Until we know more about the possible interrelationships of these variables . . . we should be cautious in interpreting cloze tests."

CHAPTER 2

The Cloze Procedure with Second and Foreign Language Speakers

2.1 Cloze with foreign languages other than English

Taylor (1954a) reported a study of the cloze procedure in Korean, in which he found the random cloze to be a good predictor of the judged difficulty of three texts in Korean. However, the subjects were native speakers of that language, and so the study does not count as an investigation of the use of the cloze procedure to measure either the readability of text for non-native speakers of a language, or their comprehension abilities.

2.1.1 Carroll et al

The first such study, and one of the most important investigations was carried out by Carroll, Carton and Wilds in 1959. Their task was to investigate the feasibility of using the cloze procedure for the College Board written foreign language achievement tests. They looked at procedures which deleted both words and letters from text, and used three groups of subjects: bilingual adults (French-English and German-English), native speakers of English with English texts, and secondary school native speakers of English learning French or German. They also investigated cloze in the auditory mode, but neither this nor the letter cloze is of concern here, and the experiment with adult native speakers of English will be reported in Chapter 3.

The investigators had hoped that cloze would prove to be simpler and cheaper than traditional test construction, that it would draw upon a broad and representative sample of language habits rather

than on "specific knowledges", that it would provide a rational scale of competence from zero up to native speaker, and that it would measure accurately at the upper levels of foreign language proficiency.

Passages were taken from the Reader's Digest, and their equivalent translations in French and German. Each of the passages contained 205 words and every tenth word was deleted, to give cloze tests of 20 items each. When these tests were given to adult bilinguals of English and one other language, it was found that there was no difference in mean scores between native speakers and non-native speakers (non-native in the sense of "second language"), and thus it was concluded that the subjects must be equally bilingual, or that the cloze was not sensitive to differences in command of one language over another. Looking at the readability of the texts, however, it was found that there was a very low correlation between the rank order of passage difficulty for, say, passages in English completed by native English speakers, and the rank order of difficulty of the same passages translated into, say, French and completed by native French speakers. They conclude that during translation the relative difficulty of text changes. Despite the finding that mean scores of the two languages for bilinguals show no difference, in fact there is only a low correlation of the bilinguals' performance in their two languages (English-French .50, English-German .06). The investigators thus suspect that the cloze procedure is not an entirely appropriate measure of foreign language proficiency, and state that "If we wish to propose cloze technique tests for measuring proficiency in a second language, it will be necessary to adjust for the individual's ability

to perform cloze tests in his native language ", since, as they suggest, "The ability to restore texts is somewhat independent of competence in a language as it is ordinarily defined."

In a subsequent section of the study, the texts used in the bilingual experiment were given to high school learners of French and German, along with the Carroll-Sapon Aptitude Test, and the College Entrance Examination Board tests. The correlations achieved with the CEEB varied greatly depending on the text used in the cloze test, and the grade level of the subjects. For example, the French cloze correlations varied from .10 to .74. A similar pattern of correlations was found for the aptitude test. The French cloze correlated at only .40 and .57 with teachers' grades for foreign language achievement in two different schools, and the German cloze correlated with similar grades at .65 and .79. The writers suggest that cloze tests are inferior as foreign language tests because they involve too much extraneous variance, and that they measure the central core of language achievement rather than some special variety of foreign language competence. This conclusion, they say, "is not inconsistent with the observation that cloze tests may not measure foreign language achievement very well."

They attempted to gain a rational measure of the amount of learning of the foreign language achieved by the school groups by expressing the mean scores for each grade level as a proportion of the mean of the adult bilinguals (resulting in, for example, the statement that fourth-year school learners have a score which is 60% of adult native speakers' performance), but unfortunately they found that this score did not show any difference in the amount of learning between

different year groups in the same school.

On an item analysis they found that there were high and low validities for both content and function word deletions, and thus concluded that little advantage is to be gained from selecting particular kinds of words only for deletion (rational cloze). They also calculated a community-of-response score, in which one point was given to an answer also supplied by at least 25% of the bilingual adults, or by 25% of the experimental group. However, these scores were just as reliable (or unreliable) as the exact word scores, correlated with them at .92, .93, and resulted in lower correlations with the aptitude test (.26 v .46). Therefore, they conclude that the extra effort involved in the calculation of such a score is not worthwhile.

Concluding their study, they do not recommend the cloze for use by the CEEB because 1) a good measure of foreign language proficiency should have native speakers performing in a relatively uniform manner, whilst learners progressively improve. This did not happen on the cloze tests; 2) the cloze seemed to be better at testing group differences rather than individual differences - that is, it discriminates between bilinguals and learners, but this is too crude a distinction; and 3) "An individual who has good mastery of a foreign language may not be able to demonstrate this mastery on a cloze procedure test if he lacks certain other intellectual qualities such as reasoning ability and ideational fluency."

2.1.2 Other studies

Apart from Carroll et al, very little has been produced to date on the use of the cloze technique with non-native speakers of

languages other than English.

Greenwald (1974) used the cloze technique to train American students of French to utilize context in their reading both of English and French, and found significant improvement in their ability to perform on a cloze test for English, but not French. She also found cloze exercises to be superior to specially constructed "contextual clue exercises" in improving cloze test performance. It does not, however, follow that cloze can be used for improving sensitivity to context clues, since exposure to cloze exercises did not result in increased scores on a specially constructed "contrived context" subtest.

Whitmer (1975) also used the cloze technique as one of a series of procedures aimed at improving American students' proficiency in reading French by exposure to special training in the strategies needed to deduce the meaning of unfamiliar vocabulary. The cloze was, however, only one part of one of four training phases (others of which were: 1) cognates and faux amis; 2) affixes, roots and word families; 3) locating main elements: subject, verb, complement, central ideas and key words; and 4) inferential techniques and cloze). The experimental group gained significantly higher scores on the Modern Language Association (MLA) Cooperative French Test (Reading) and so the author concluded that his techniques worked. He also suggested that the use of cloze as a teaching technique needs further investigation, and that cloze shows promise as a device for testing proficiency in a second language.

The only recent study validating the cloze as such a test for foreign languages other than English known to the author, however, is

Binkley (1974), which assessed cloze both as a teaching technique and as a measure of German reading comprehension. She gave a total of six cloze tests based on German passages at two-week intervals. Each cloze pre-test was followed by study of the intact text and by discussions of cloze, used to teach structure and vocabulary. The cloze post-test was given ten days later. Thus twelve cloze scores, as well as a total cloze score, were available for correlation with the MLA Cooperative German tests, coefficients for which ranged from .59 to .89. Higher correlations were achieved by combining pre- and post-test scores. Since the highest coefficients were obtained between cloze and whichever MLA test was appropriate for the students' achievement level, the author concluded that cloze is more flexible than the MLA test, which must be adjusted to student level. Interestingly, no significant correlations were obtained between the cloze and a test of English reading ability - a finding on which the author makes no comment, but which suggests that the ability to do cloze in German at least is not related to one's reading abilities in one's native language. Binkley advocated the use of cloze as a classroom measure, especially for ongoing, cumulative evaluation, but pointed out that further research is necessary to determine the extent to which German cloze reflects students' full linguistic capacity.

2.2 Non-native speakers of English

With non-native speakers of English, however, somewhat more work has been done. One of the first studies was carried out by Friedman (1964), whose subjects were foreign students at the University of Florida. She used the cloze procedure not as a testing device,

but as a teaching technique for improving her students' reading comprehension. She found, however, that doing cloze exercises did not improve comprehension as measured by a traditional multiple-choice test more than did a course in close reading. She did note that her students reported they were more aware of the grammatical structures in text when doing the cloze.

Levine (1971) also used cloze in a teaching situation with students of English as a Second Language, but this time the cloze was the testing device used to measure learning hypothesised to have taken place during an audio-visual, conversational course. She found that there was no significant increase in cloze scores after instruction. She maintained, however, that the cloze was an effective measuring device of the progress by students in courses for foreign language learning, and thus concluded that the audio-visual course in English caused students to neglect their grammar and to concentrate on pronunciation. The other possible conclusion, which Levine did not consider, is that learning did take place but that the cloze tests were not sensitive to it.

Two studies have been made on the use of cloze tests as tests of listening comprehension for foreign students of English. Gregory-Panopoulos (1966) deleted every fifth word from an aural text and found that the ability of students to restore words correlated closely with a standard listening comprehension test. Moreover, he found that the reliability of the cloze test was higher. Templeton (1973) also used the cloze procedure to test the aural proficiency of foreign students, found that it correlated well with criterion English

proficiency tests, and suggested that it could be used to identify weak students amongst those foreign students coming to study in Britain. Interestingly, in his pilot study, Templeton found that the frequency of the test items significantly affected the subjects' performance.

2.2.1 Readability

Several studies have been concerned with the use of the cloze procedure to measure the readability of text for students of English as a second or foreign language. Anderson (1972) used non-native speaker primary school subjects in Papua, New Guinea, and administered cloze tests where every 8th word had been deleted from three separate passages. Four experienced teachers, involved with children's reading, ranked the passages in order of difficulty. The cloze ranked the passages in exactly the same order as the judges, regardless of how the cloze tests were scored. He concluded that cloze was a valid measure of readability for non-native speakers of English.

Oller (1972) also examined the cloze as a measure of readability, this time with foreign students at the University of California at Los Angeles. He took three texts, which varied in difficulty according to both the Flesch and the Dale-Chall readability formulae, and deleted every 7^{th} word from the passages. He found that although the cloze test scored for exact word only agreed with the formulae's prediction of difficulty, the "any acceptable word" scoring procedure resulted in a different order of difficulty of texts from that predicted. Apart from speculating that the difference in rank orders

might have been due to a difference in linguistic redundancy in the texts not detected by the formulae, Oller does not discuss this finding, nor does he question the suitability of the readability formulae for use with EFL students (unlike Anderson (1972) who, regarding the formulae as inappropriate for such students, used judges to establish text difficulty).

Haskell (1973) used six texts, three of which were easy and three of which were difficult, as determined by reference to the Thorndike Word List (texts with words at the 500 word level were regarded as easy, those with words at the 1,500 - 3,500 word level were called difficult). He found that a discrimination between easy and difficult passages was achieved regardless of whether the texts were fiction or non-fiction, whether every 5^{th} or every 10^{th} word was deleted, whether 50 words were deleted or only 35, and whether the blanks replacing the deleted words were of uniform size or whether they varied according to the size of the deleted word. Unlike the Oller study, he found that the scoring procedure used had no effect on the rank order of difficulty - the discrimination between easy and difficult was maintained when only the exact word was scored, when synonyms were allowed, and even when "any appropriate word" was counted as correct.

2.2.2 Cloze as a measure of linguistic proficiency

With non-native speakers of English, the cloze has been compared not so much with tests of reading comprehension as with tests of proficiency in English as a second language. For example, Johnson (1974), using Air Force officers as subjects, concluded that cloze was

an effective measure of ESL, because it was significantly related to other English proficiency tests. (He also found, as had been found previously with native speaker subjects, that deleting only structural words resulted in a significantly easier test than an every-n^{th}-word deletion, and deleting only "lexical" words resulted in a more difficult test.)

The first study of this nature was carried out by Darnell (1968). His scoring procedure was somewhat idiosyncratic, in that it involved comparing the responses of non-native speakers with those given by native speaker subjects, rather than with some external arbitrary criterion of correctness. In his pilot study, Darnell found that his clozentropy scores (based on cloze tests where every 5^{th} word had been deleted) correlated .63 with a dictation test, .61 with the Gates Reading Survey, but not at all with an oral interview. He viewed these results as encouraging for further research with the clozentropy, and so he carried out a validation study of the procedure based on four passages, two on engineering and two on a liberal arts subject, two easy and two difficult. He turned these texts into cloze tests by deleting every 10^{th} word and compared them with the Test of English as a Foreign Language (TOEFL). The total cloze score, based on all four passages, correlated .84 with the TOEFL total, and it also correlated highly with the Listening Comprehension (.74), English Structure (.67), Vocabulary (.73) and Writing Ability (.70) subtests. Interestingly, the lowest part-score correlation was achieved with the Reading Comprehension subtest (.60).

Since, in addition, the reliabilities of the TOEFL and the

total cloze scores (based on 200 items) were virtually identical (.86), Darnell concluded that the two tests were measuring, for all practical purposes, the same thing, and that, to the extent that the TOEFL is an acceptable measure of English proficiency, the clozentropy battery must also be acceptable. It should be noted, however, that different cloze tests in the battery correlated differently with the TOEFL, with coefficients ranging from .49 to .70 with the subtests for different cloze tests, and from .63 to .77 with the total TOEFL score.

Unlike most other researchers, Anderson (1970) compared his cloze tests with tests of reading comprehension, specifically the Watt's Reading Comprehension Test (intended for native speakers) and a specially made multiple-choice test on the same passages as the cloze tests, constructed to test six comprehension skills (knowledge of the vocabulary used in the passages, knowledge of stated facts, ability to perceive sequences of events, to see relationships, to identify the main theme, and to make inferences) in an item ratio of 36 to 18 to 9, 9,9 and 9.

He used nine passages containing a total of 50 items, where the items were selected (from cloze tests administered to competent foreign readers) for the unanimity with which subjects had provided responses. In other words, his cloze tests were not random, but rational. His subjects were primary school learners of English as a foreign language in Papua, New Guinea. His cloze tests proved to have individual reliabilities of .8 and above, and a total reliability of .95. He found a correlation between total cloze score and general reading comprehension (the Watts test) of .78, and between cloze and

specific comprehension - the homemade test - of .85. All intercorrelations were significant: the different cloze passages correlated with each other around the .85 level, and with the Watts test and the vocabulary section of the homemade tests, at between .71 and .78. However, the correlation of the cloze with the subsections on organisation, main theme and inference were noticeably lower, ranging from .46 to .55.

Anderson also carried out a factor analysis of his results (the only such study to be done with non-native speakers) and found that only one factor, which accounted for 65% of the variance, emerged. The cloze tests had the highest loadings on this factor, which can only be described as "reading comprehension". He therefore concluded that the cloze test is a highly reliable measuring instrument and a valid measure of both specific and general reading comprehension.

Most of the work on cloze with non-native speakers has been carried out by John Oller and his associates, and has concerned itself with cloze as a measure of proficiency in ESL. The first such study was done with Christine Conrad (Oller and Conrad, 1971), where it was found that the cloze correlated with the total score on the UCLA English proficiency test at .88 (by multiple regression). The highest correlation of cloze test with subtests of the UCLA test was with the dictation test, at .82, and the next highest was with reading at .80. Much lower correlations were achieved with tests of vocabulary (.59), grammar (.58) and the article (.33). Foreign students studying at UCLA were the subjects, and they were divided into five groups: beginners,

intermediates, advanced, a composition group, and graduates. Their performance on a cloze test where every 7th word had been deleted was compared with that of two groups of native English speakers: college freshmen and college graduates. When the exact word method was used to score the cloze tests, it was found that mean scores for the first two non-native groups were significantly different from each other, and from the seventh group (graduate native speakers). However, the cloze test failed to discriminate among the advanced, composition and graduate non-native speakers, or any of these with the *freshman native* speakers. If it is assumed that the native speakers, whether freshmen or graduates, are homogeneous, and different from non-native speakers - a normal, pragmatic assumption - then it appears that the cloze is not only discriminating falsely among native speakers, but also failing to discriminate where it should - between native and non-native speakers. The authors seek to explain their finding in two ways: either the native speakers are in fact heterogeneous and advanced non-natives are indeed similar to freshmen native speakers, or the scoring procedure used is inappropriate and an "any acceptable word" method might produce more intuitively satisfying results. Clearly more research on this matter is needed.

Oller and Inal (1971) constructed a rational cloze test which deleted only prepositions, and which was scored by the "any acceptable word" method. Significant differences were obtained between mean scores for native speakers and for non-native speakers. Moreover, the correlation of item difficulty for native speakers and non-native speakers was merely .23. Somewhat contradicting the assumptions of Oller and

Conrad (1971), the authors suggest that, assuming that all normal literate adult native speakers have roughly comparable skill in the use of prepositions, and since native speakers were not discriminated by this test, "this cloze test of prepositions is actually a test of English language proficiency, rather than of some other language-related skill on which native speakers might be expected to differ significantly." In other words, a language proficiency test should not discriminate among native speakers, should separate them from non-native speakers and should discriminate among non-native speakers. If one relates this to the Oller and Conrad study, one is forced to ask whether their cloze test (every-n^{th}-word) is a language proficiency test.

In the Oller and Inal study, the results of one group with different linguistic backgrounds were correlated with their performance on the UCLA English proficiency test and reasonably high coefficients were achieved with all the test subsections (.63 - .69). The correlation with the total test score was .75. Interestingly, when partial correlations were carried out, the highest coefficient was with grammar. Vocabulary, dictation and reading were all non-significant. The authors thus conclude that their rational test is a valid test of grammatical competence.

Oller (1972), as is reported in more detail in Chapter 3, found the "any acceptable word" scoring procedure to be capable of better discrimination among subjects than the exact word procedure, and of higher validity correlation coefficients with one version of the UCLA proficiency test. Correlation with the total test was .83 and the highest part correlation was, again, with dictation (.80), followed by

reading (.76), grammar (.72) and vocabulary (.64). This order of relation was obtained even when subtests were partialled out in the equations. Oller explained his findings by reference to the relative integration of the tasks on the subtests, such that, as the tasks became more integrative, their correlation with the cloze increased. The vocabulary test, which merely requires synonym matching, is not at all integrative, according to Oller, and so correlated the lowest. Grammar requires what Oller termed syntactic integration, and thus correlated higher than the vocabulary test with the cloze. The dictation, being a more complex task than the reading test - it "demands more active hypothesis testing and analysis by synthesis than does the reading test" - correlates highest of all. Thus, Oller concludes that cloze also requires high-level integrative skills.

Oller, Bowen, Dien and Mason (1972) deleted every 6th word from seven texts, some in English, some in Thai and Vietnamese, where the texts in differing languages were translation equivalents of one another. They intended to compare native and non-native performance on cloze tests, in particular, to determine whether native and non-native responses differ. Scoring was done by the "any acceptable response" method. When examining the mean scores of the original passages and their translation equivalents clozed by native speakers, an approximate equivalence was discovered. The authors conclude that translating a cloze test into another language, if done carefully, yields a cloze test with a level of difficulty comparable to the original. (This is contrary to the findings of Carroll et al, 1959.) Moreover, however, the authors suggest a pedagogic application of this finding:

"Suppose you are a teacher of a couple of advanced Russian classes of about equal proficiency in Russian, where the students are native speakers of English. You want to know how your students compare in ability with native speakers of Russian of similar socio-economic status and educational background. Here is a fairly simple way to make an approximate judgement. Carefully translate a Russian passage into English maintaining comparable style level, etc. Make a cloze test of it by deleting, say, every 6^{th} word. Give the test to one of your Russian classes. The mean score will tell you approximately what native speakers of Russian would score on the original passage in Russian. Give the original cloze passage in Russian to your other class. Their mean score subtracted from the mean score of the first class on the English passage should provide a global but useful indication of their competence in Russian relative to native speakers of Russian."

The authors thus suggest the equivalence of the cloze task across languages, and propose that this knowledge can be used to construct objectively criterion-referenced language proficiency tests.

They also found that whilst non-native speakers frequently made responses to the cloze task which violated some contextual constraint, native speakers rarely made this kind of response. In particular, non-native speakers made considerably more responses which violated obligatory selectional constraints in the immediate phrase structure, and which made no sense at all. The authors thus hypothesise that non-native speaker responses tend to be different in type as well as quantity from native speaker responses.

Oller, Atai and Irvine (1974) presented a more traditional study of the cloze test as a measuring tool for second language proficiency, by correlating a test on a passage where every 7^{th} word had been deleted, with scores on two dictation tests and the TOEFL test. The latter test correlated .78 with the cloze, but the highest part-TOEFL – cloze correlation was with the Listening Comprehension test (.76). This replicates Darnell's (1968) and Oller and Conrad's (1971) findings. Correlations with dictation were only slightly lower, at .69, and were higher than with any other part-test of the TOEFL (reading comprehension, .67; structure and writing ability .66; vocabulary, .49). The authors thus conclude that test modality has little effect on the results when what is being measured taps a source common to the written and spoken modes, namely, what they call the learner's "internalized expectancy grammar", or his underlying language competence. The reason that cloze, dictation and the listening comprehension test tap this competence, they claim, is that they are all integrative rather than discrete-point tests, they are task-oriented and require the pragmatic use of language for communicative purposes. Since these three tests were more highly correlated with one another and with the other part tests of the TOEFL than the latter were with each other, the authors conclude that integrative testing procedures are more valid than other procedures.

They also compared cloze scored by the exact word method with the same test scored by allowing responses to be correct if they fit all of the surrounding context. Since no mean scores nor standard deviations were given, it is difficult to evaluate the second scoring

procedure in terms of how much additional information was provided, but the intercorrelation of the procedures was .94, and the correlations with TOEFL total and part-tests were virtually identical. The only difference between the two was that the acceptable-word method correlated higher with dictation (.75 v .69). Given the thesis that integrative tests will intercorrelate more with each other than with other procedures, one would have expected the authors to have concluded that the acceptable word procedure provided a more integrative test, but in fact they conclude by recommending the exact word procedure for use by non-native speakers, teachers of ESL, whose judgment of "acceptability" is assumed to be less valid and reliable than that of native speakers. This issue has not been put to the test.

Finally, Stubbs and Tucker (1974) also investigated the relationship between the two most common cloze test scoring procedures and, finding that there was a correlation of .97 between the exact word and the any-acceptable-word procedures, recommended use of the former as the simplest and most reliable. The cloze test, constructed by deleting every 5^{th} word from text, was correlated with the American University of Beirut Entrance Test of proficiency in English, and since moderately high coefficients were achieved, the authors conclude that the cloze is a valid test of English proficiency. However, the "acceptable word" scores consistently correlated higher with the AUB test and its sections than the exact word scores (coefficients with the total were .76 and .70 respectively; with structure, .70 and .66; vocabulary, .65 and .60; and with reading, .70 and .67). It would thus seem, yet again, that one might just as easily have concluded that the acceptable word

procedure is in fact the most valid, and hence the procedure to be recommended.

2.3 Summary

Since there have been far fewer studies of the cloze procedure with non-native speakers of a language, and very few comparisons of the performance of native and non-native speakers on the same test, there is little evidence for the validity of the procedure used with non-native speakers, or for the similarity or difference between native and non-native performances. Some of the evidence produced is inconclusive; in particular, as with native speakers, the use of the cloze as a teaching technique seems to offer no advantage over other teaching techniques, and there is some doubt as to the ability of the cloze to measure learning in a foreign language.

The evidence available suggests that cloze might be a suitable measure of text readability at least to the extent that experienced teachers and readability formulae provide suitable estimates of text difficulty.

Some evidence is also available to suggest that the cloze is a valid measure of foreign language proficiency, since it correlates well with more traditional, and presumably valid, measures of such proficiency. A strange and fairly consistent finding is that the cloze tends to correlate higher with tests of listening comprehension including dictation, than it does with tests of grammar or reading comprehension. Whereas high correlations are obtained with reading tests intended for native speakers, lower correlations are apparent with those

subtests of such reading tests which require higher-order reading skills. Notwithstanding the correlations with linguistic proficiency tests, there is some doubt about the ability of cloze to test specifically foreign language achievement. Carroll et al claim that a good test of such achievement should result in native speakers performing uniformly well, and non-native speakers being discriminated. There is some evidence that the cloze test does not do this, which leads to a suspicion that ability to do cloze may be different from competence in a language.

The validity coefficients for the cloze seem to vary according to the difficulty of the test used, and also according to the scoring procedure employed. There is as yet no consistent use of one procedure over another, despite high intercorrelations between procedures, because of this variability of validating correlations.

Little work has been done with the rational cloze procedure, but with the random procedure there seems to be an assumption that deletion frequency has no effect, since different investigators use different frequencies.

In summary, then, it does seem that the cloze procedure is a potentially interesting measure of language proficiency for non-native speakers and of text readability for the same subjects, but that the influence of different variables in the cloze technique needs to be investigated more closely, in an attempt to reveal what it is that the cloze test might be testing.

CHAPTER 3

Some Aspects of the Cloze Procedure

As we have seen in previous chapters, the cloze procedure is a complex phenomenon, about which it is misleading to make global statements. Although not a great deal of attention has been given to the effect of different variables in the past, some evidence has been presented to suggest that changing the value of some of these variables may well have an effect on the validity of the cloze by changing its nature. This chapter looks in more detail at previous research into the effect of these variables.

3.1 Pre-cloze tests and post-cloze tests

A pre-cloze test is a cloze test taken by a subject without reading the unmutilated text. A post-cloze test is a cloze test taken after first reading the unmutilated passage. The latter form is somewhat more time-consuming and awkward to administer, and so the usual cloze procedure has been the pre-cloze.

Taylor (1957) found slightly higher correlations between his post-cloze test and his comprehension test than he did between the same comprehension test and the pre-cloze test. His correlations with cloze tests and the Air Force Qualification Test also tended to be somewhat higher for post-cloze than for pre-cloze. However, the differences were almost certainly not significant.

Rankin (1957) found similarly higher correlations between cloze tests and the Diagnostic Reading Test (Survey Section) for the

post-cloze, but only one difference in coefficients was statistically significant. However, he also found that the pre-cloze validity and reliability both vary as a function of the subject's personality, whereas those of the post-cloze do not. He therefore concludes that whilst pre-cloze is suitable for the measurement of groups of subjects, the post-cloze would be more appropriate for the study of individuals. Bormuth and MacDonald (1965), however, found no difference in validity coefficients between pre- and post-cloze tests when investigating cloze as a measure of sensitivity to style.

Although Rankin (1957) found a significant increase in post-cloze scores over pre-cloze when looking at reading gain, his deletion system was rational, rather than random. Coleman and Miller (see Chapter 1) concluded that pre- and post- cloze tests were measuring the same thing. Interestingly, Greene (1964) found no significant difference in mean scores between subjects taking a cloze test and subjects who read the intact passage before taking the cloze test. This finding suggests either that no learning takes place when reading a passage before taking a random cloze test - which seems unlikely - or that reading a random cloze test gives one the same information, and is the same sort of task, as reading unmutilated text - a conclusion which also seems counter-intuitive. Perhaps more likely is the assumption that the random close test does not test the sort of comprehension one acquires from reading unmutilated text. This theme will be taken up again in section 3.7 on the sentence-bound nature of cloze.

No investigation of pre- and post- cloze differences has been carried out using non-native speakers as subjects.

3.2 Rational cloze and random cloze

As mentioned in Chapter 0, pseudo-random cloze is the type of deletion most frequently used, rather than pure random deletion. Taylor (1953) compared an every-10^{th}-word deletion with a random 10% deletion and came to the conclusion that, provided more than 16 items were used in the cloze test, there was no difference between the procedures, and that an every-n^{th} deletion was to be preferred for convenience. In the comparison of rational and random cloze, therefore, the random cloze used is in fact a pseudo-random procedure.

Taylor (1957) compared a pseudo-random cloze with two rational deletions: 1) the deletion of easy words only (conjunctions, pronouns, articles, auxiliaries), and 2) the deletion of hard words only (nouns, verbs, adverbs). With only one exception, the random cloze correlated best with criterion tests of pre-reading knowledge, immediate recall and aptitude, and so he concluded that for purposes of testing comprehension, aptitude and readability, the random procedure was best.

As already mentioned in section 1.5 of Chapter 1, Rankin (1957) found the rational cloze to be less related to intelligence measures, and more to "pure" comprehension of the text, than was random cloze.

Greene (1964) found increased validity coefficients for a modified cloze procedure on some criteria, but not on others, and concluded, despite better item discrimination and reliability for his modified cloze, that there was no significant difference between the two deletion types as measures of reading comprehension.

Jefferson (1969) investigated two rational procedures - deleting every 5th lexical unit (noun, verb, adjective) and deleting every 5th structural unit (function words) - and compared them with a random every-5th-word cloze. He found significant differences in mean scores, such that lexical cloze was hardest, followed by random cloze, and a significant interaction was found between deletion type and readability assessment. He claimed, therefore, that using rational rather than random cloze seriously affects the resulting measures of readability. This is the same conclusion reached by Taylor in 1953, who saw that if the texts whose readability one is measuring are to be sampled adequately, then random cloze is the only possible procedure, since the rational cloze would result in a biased view of text difficulty.

Doyle (1973) compared a rational lexical deletion (deleting every 10th noun, verb, adjective or adverb) with a random every-10th-word deletion. He discovered that his subjects performed more successfully on rational cloze for expository text, but more successfully on random cloze for narrative text. He claimed that rational cloze requires teleological processes, measures reading achievement, and samples linguistic deep structures, whereas random cloze measures the ability to comprehend interrelationships among ideas, and samples linguistic surface structures. He did not find IQ to be more associated with one procedure than the other.

Prange (1973) found no significant difference between random (every 5th word) and rational (every 3rd noun, verb and adverb) procedures and correlations with critical reading, general reading and intelligence measures.

Rankin (1974), reviewing cloze research over the previous twenty years, claims that the concentration on random cloze procedure has "strengthened the influence of general verbal abilities and intelligence upon the cloze measurement of reading comprehension." He advocates further investigation of rational deletion procedures, but wonders whether such rational procedures can then be considered to be cloze procedures, whose essence, in Taylor's days, was the random sampling of linguistic items in text.

Despite some *counter-evidence*, *the consensus* of opinion on rational cloze tests with native speakers seems to be that they are different from random procedures, and, of course, that they are capable of greater manipulation and variation. They may, therefore, be more suitable for investigating the effect of linguistic variables on comprehension than are random procedures.

To date, the only investigation known to this author into rational cloze procedures with non-native speakers of English, or indeed of any other language, is the Oller and Inal study (1971), reported in Chapter 2, where only prepositions were deleted from text. The test was administered to non-native speakers of English, and the results were compared with the UCLA English proficiency test. Although, as reported, the authors claimed theirs was a test of grammatical competence, no comparison was made with a random deletion procedure on the same text in order to establish whether the rational test was testing anything different.

3.3 Cloze as a test

Not many investigators have looked at the efficiency of cloze tests, but of those that have, opinions have generally been favourable.

Greene (1965) reports that his modified cloze procedure, deleting only those words he considered to be restorable, resulted in a better reliability coefficient (KR21 .76 v .9, split half .5 v .76) and superior item performance. The mean item difficulty for modified cloze was .58; for random cloze the content word difficulty was .38. More effective items were present in the modified cloze, for the random cloze had many items that did not discriminate and 13% of the items were extremely difficult.

Bormuth (1965a) reported that the frequency distribution of item difficulties tends to be U-shaped, which echoes Greene's finding. However, Cranney in Greene (1967) claims that rejecting items after item analysis does not improve the correlation with validating criteria, and that, in fact, it lowers reliability. This lower reliability, of course, is due to the smaller number of items in the test. Reliabilities of cloze tests, when reported, have tended to be moderately high (of the order of .7 and over), but only when sufficient items have been included. Taylor (1953) felt that acceptable reliability would be achieved if more than 16 items (and preferably at least 35 items) were included in the test, and Bormuth (1962, 1963) recommends that 50 items should be included for optimum reliability. This, of course, depends also on the number of subjects, and in fact Bormuth (1965a) presents a table of varying standard errors of the mean for various combinations of test length and numbers of subjects, for the guidance of researchers.

However, both Taylor (1954b) and Bormuth (1964c) found that, if, using the every-5th-word deletion rate, five versions of the cloze test on a particular passage were constructed so as to delete every word in the passage, even if 50 words were deleted, more than half of the deletion versions were significantly different. Also with native speakers, Carroll et al (1959) found that when holding ability to do cloze constant, there was a significant difference between a cloze test with a deletion rate of every 10th word, starting at the tenth word, and one of the same deletion frequency, but starting at the eleventh word.

In summary, it seems that, although reliabilities may be high if sufficient items are included, many of the items in a random cloze are in fact contributing very little to the test as a whole, and so the test can be considered to be fairly inefficient. Yet item analysis does not improve validity, and may affect reliability.

The only investigation of cloze efficiency with non-native speakers known to the author was carried out by Oller (1972), using non-native speakers of English as subjects. Comparing the exact word scoring procedure with the any-contextually-acceptable word procedure, on texts of varying difficulty, Oller found that although reliabilities measured by KR 20 were high (from .93 to .99), the reliabilities for items improved slightly when the acceptable-word procedure was used. He found that item discrimination was worse when the exact word method was used, regardless of text difficulty. On the difficult text, 16% of the items failed to discriminate using the exact word method, and 18% were extremely difficult. On the medium text, 38% of items were very easy (above 80% facility) and 36% on the easy text. Moreover, 34% of

the easy text items failed to discriminate between "good" and "bad" pupils. In other words, there was a very uneven distribution of item difficulties and discriminations with the exact word scoring method. Whilst using an any-acceptable-word procedure improved discrimination somewhat, it also resulted in a much more unbalanced distribution of item difficulty (the easy text proportion of items with an item facility of over 80% increased from 36% to 80%, and the difficult text increased from 4% to 25%). There are grounds for doubting the efficiency of the cloze test with non-native speakers, at least when viewed from the point of view of traditional item analysis - a technique intended for use with discrete-point tests, but also used to analyse so-called integrative tests.

3.4 The effect of passage difficulty on cloze scores

As has already been seen, the cloze procedure is generally considered to be sensitive to differences in the difficulty of texts, although Mosberg et al (1968) expressed doubts as to the sensitivity of cloze to passage differences at low levels of difficulty. The problem that presents itself is whether the text used for a cloze test affects the measurement of the subject's comprehension of the text or of his general reading comprehension abilities. Clearly, different raw scores will result for each individual on different texts, which is, of course, the justification for claiming that cloze measures readability. But perhaps individuals will differ as to the difficulty they find on different texts, and, therefore, the correlations with other comprehension tests will change.

No such study known to the author has systematically compared the way in which cloze tests on texts of differing difficulty measure comprehension abilities differently, for native speakers. However, one or two studies have revealed information about the effect of text difficulty on the measurement of comprehension for non-native speakers. Carroll et al (1959) found that the correlation of their cloze tests with the CEEB varied as the text used for the test changed in difficulty.

Darnell (1968) compared four cloze tests, two difficult and two easy, two on engineering and two on the humanities, with the TOEFL test. The scores he used were clozentropy scores (a type of communality-of-response score, using the frequency of different responses of a criterion group of native speakers as the scoring key for non-native speakers) and so direct comparison with normal exact word scoring is no simple matter. Differences in the correlations with the TOEFL were found between easy and difficult texts (difficulty as measured by Flesch), but they were minimal (Easy .69, Difficult .63 for engineering texts; Easy .73, Difficult .77 for non-engineering texts). The greater differences were found between texts of different <u>content</u>, such that non-engineering texts always correlated higher with the TOEFL than did engineering texts (the subjects were engineering students). To the extent that difference in content reflects a difference in text difficulty for these subjects, even if the difference is not measured by Flesch, it can be shown that text difficulty does affect the measurement of language proficiency.

Oller (1972) used three passages (all much easier than Darnell's passages) which he called "very easy", "fairly easy" and

"standard" (Flesch 100, 77, 69 respectively). Interestingly, his "any acceptable word" scoring method resulted in a reversal of order of difficulty of Texts I and II, but even using the exact word, the difference between fairly and very easy texts was not so great (mean scores 64% and 68%). Nevertheless, the very easy text, Cloze I, consistently correlated considerably lower with the criterion UCLA ESLPE English proficiency test than either of the other two texts. (The correlation with the total UCLA test was .73, compared with .87 for Cloze II and .85 for Cloze III). Between the fairly easy and the standard texts there was very little difference. This finding tends to confirm the finding of Darnell, that text difficulty might have an effect on the validity of cloze. Since, however, Oller's texts were only subjected to one deletion each (every seventh word), it is possible that the choice of words for deletion had an effect. Further investigation is needed with several deletion patterns per text, and with texts which are more obviously different in difficulty.

3.5 The effect of the scoring method on cloze scores

Whilst emphasising the need for further research, Taylor's original paper (1953) claimed that scoring correct only the exact replacement of the word deleted yielded the same degree of differentiation between scores and passages as a scoring method which allowed as correct any synonym (also defined as a "good enough" answer), although the second method gave a slightly higher score. He concluded that exact replacement was to be preferred on practical grounds.

Rankin (1957), Ruddell (1964) and Bormuth (1965b) confirmed this finding. Rankin found a correlation of .92 between exact and

synonym scores on a pre-cloze test, and found that there was no significant difference in reliability between the two scores. Ruddell's synonym score judged responses correct when they 1) completed the original idea expressed in the context of the sentence, 2) fitted the original syntactic pattern of the sentence, 3) were grammatically correct in terms of number agreement, and 4) corresponded in difficulty level (judged by absence from the Dale 3,000-word list) with the deleted word. (It should be noted that this is a different synonym score from that used by others, who simply used a dictionary of synonyms to decide on the acceptability of responses.) Ruddell found no significant difference in reliability between the exact and the synonym scores, although there was a tendency for the synonym reliability to be higher. Nor did he find any significant difference between exact and synonym scores when correlating them with the Paragraph Meaning section of the Stanford Achievement Test (although, again, there was a slight tendency for higher validity coefficients for the synonym scores).

Bormuth (1965b) classified responses to 20 cloze tests into seven different categories: 1) exact word, grammatically correct (EGC); 2) exact word, grammatically incorrect (EGI); 3) synonym, grammatically correct (SGC); 4) synonym, grammatically incorrect (SGI); 5) unrelated semantically, grammatically correct (USC); 6) unrelated semantically, grammatically incorrect (USI); and 7) unclassifiable responses (UCR). (Bormuth's definition of synonym is unclear, but seems to be opposed to unrelated semantically; otherwise, there is a great difference between scores 3 and 5.) He found that while scores based on gramma-

tically correct responses correlated positively with his criterion of comprehension (the Stanford Achievement Test), the scores based on grammatically incorrect responses did not. Moreover, the correlations of grammatically correct responses with comprehension increased as a function of the similarity of the meaning of the responses to the deleted word (correlations of the Stanford test with UGC, .55; with SGC, .64; with EGC, .82). This he took to mean that comprehension could not be said to be complete unless the exact word had been replaced, and so he concluded that the exact word score was the best, both for measuring individual differences in reading ability and for discriminating among passages.

Bormuth (1968b), reviewing studies of different scoring procedures, claims that when higher validity coefficients are achieved with the synonym scoring method, this is simply because the variability of the scores has increased. Thus, one can compensate for this by adding a few items to the test, and then using the exact word procedure.

Coleman and Miller (1967) compared the exact word score with a weighted score which gave a value of 3 to the exact word, 2 to a synonym of the deleted word, and 1 to a word of the same form class as the deletion. Possibly because of this weighting, the intercorrelation of the two procedures was .99.

Although many studies report high intercorrelations between the exact word score and other scoring procedures, therefore concluding that the different procedures either all measure the same thing or at least that there is no need to calculate other scores, Hafner (1964) found a relatively low correlation (.61) between the exact word score

and a "grammatically correct" score, as he called it (in fact, a sort of form class score, since correctness was determined by reference to the form class of the deleted word), which he assumed to measure sensitivity to immediately adjacent context. He also found that his grammatical score correlated less with intelligence than did the exact word score (Otis IQ Test .68 versus .73, Hafner Intelligence Test .42 versus .46) and less with a measure of vocabulary (non-significant versus .56). He did not reach any conclusions as to the usefulness of his grammatical score. Nevertheless, his results show that some scoring procedures may well measure different aspects of reading comprehension. Indeed, it would be surprising if it were not so.

Fillenbaum, Jones and Rapoport (1963) also employed a form class scoring procedure - i.e., giving credit for any restoration from the same form class as the deleted word - but they did not compare it with the exact scoring procedure, nor with external criteria, since they were concerned solely with the predictability of form classes. Nonetheless, they make the interesting suggestion that possibly a form class scoring procedure measures sensitivity to the relatively close grammatical environment, whereas the exact word score might depend more on remote semantic features of the discourse. They did not, however, put this hypothesis to the test.

Moreover, despite the apparent evidence that the exact word procedure is perfectly valid and reliable, some investigators have used other procedures because they <u>felt</u> the exact word method was too harsh on subjects. For example, Schoelles (1971) decided that with children in grades 1 - 5, it would be more appropriate to allow synonyms of the

deletion as well as the exact replacement.

Nevertheless the usual practice remains, with native speakers, to score only the exact word. With non-native speakers, however, there has been a feeling that an exact word scoring criterion may simply be too difficult, and that a better measure of proficiency might be to allow words which are contextually acceptable. Take the sentence "The ..... walked down the street". If a subject responds with "horse", "bear" or "bird" instead of "woman", the error would seem to be of a different order from filling the gap with "table" or "with".

One procedure which has been examined several times is the communality-of-response score, where native speakers fill in the cloze test, and non-native speakers' performances are judged by the words supplied by the (supposedly proficient) native speakers. Carroll et al (1959) gave non-native speakers a credit point for any answer also given by either 25% of the native speaker subjects, or by 25% of the non-natives themselves. However, they found that the reliability of this score was just the same as for the exact score, and that its validity was less - the exact word correlated .46 with an aptitude test, whereas the communality score correlated only .26. (The intercorrelation of the two scores was .92.) They thus concluded that the exact word score was a better measure of language achievement.

Darnell (1968) used a clozentropy score, based on logarithms of the frequency with which criterion groups of native speakers had responded to the same cloze tests, and achieved encouragingly high validity correlations with the TOEFL (for example, total clozentropy with total TOEFL was .84). He therefore concluded that the clozentropy

procedure was a valid measure of proficiency in English as a foreign language. Unfortunately, however, he did not compare the clozentropy with the exact word score, and was thus unable to comment on the relative merits of the two procedures.

Levine (1971), like Hafner and Fillenbaum et al, supposed that scoring cloze passages for correct form class replacement would result in a score expressing the subject's demonstration of his English grammatical competency, but did not supply any proof of the assertion. She did, however, find that whereas the cloze exact score did not change after a course of instruction in English (test given before and after the course), the form class scores did. Unfortunately the "after" scores were lower than the "before" scores.

Anderson (1972) examined four scoring procedures: 1) the exact word, 2) synonyms (those words appearing in the Collins Gem Dictionary of Synonyms and Antonyms as synonymous with the deleted word), 3) alternative responses (defined as those words which "made sense within the context, which were correct in number agreement, and which fitted into the grammatical structure".), and 4) same grammatical class as the deletion. Unfortunately, scores 2, 3, and 4 were weighted in favour of the exact word by giving at least twice as much credit for verbatim restoration as for any of the others. Therefore, the intercorrelations of the four methods were high (none was below .99) and the reliabilities were virtually identical. Since, although the mean scores were higher on methods 3 and 4, the texts were ranked in the same way by all four methods, and since virtually maximum intercorrelations of the procedures were achieved, Anderson decided that the exact word method

was the most appropriate for testing purposes.

Oller (1972) examined five scoring procedures with 398 students of English as a foreign language. His procedures involved different weights for different categories of responses: M1: exact word only; M2: exact words and any other contextually acceptable responses - i.e., restorations that violated no contextual constraints. M3, M4 and M5 involved different weightings for exact words, acceptable responses, responses which violated long-range constraints, and responses that violated short-range constraints, in the following proportions: M3: 4 + 3 + 2 + 1; M4: 2 + 2 + 2 + 1; M5: 2 + 2 + 1 + 1 . Since he found that M3, 4 and 5 were not significantly different from M2 (not altogether surprisingly, in view of the weighting systems) and were much more complex, he rejected them and concentrated his analysis on M1 and M2. He discovered that validating correlations with an English proficiency test were consistently (regardless of the difficulty of text used for the cloze test) higher for the acceptable-word scoring method, with the sole exception of correlations with the vocabulary subtest. (Overall correlations of cloze with the UCLA proficiency test were exact word, .75; acceptable word, .83.) As reported in section 3.3 above, item analysis revealed greater discrimination for M2, but an increase in item facility. Oller therefore concluded that with non-native speakers, the best scoring procedure is that which allows any contextually acceptable word as well as the exact word.

The following year Haskell (1973) reported a study of the cloze for measuring readability of text rather than comprehension abilities, among non-native speaker subjects, in which, inter alia, he

compared three scoring procedures: exact word, synonyms, or any-appropriate-word. He found that all three methods ranked his "easy" and "difficult" texts in exactly the same way, and that the increase in mean scores for each procedure was quite small (7% increase of synonyms over exact word, 5.5% increase of any-appropriate-word over synonyms). Thus he decided that the exact word method was the most suitable measure of readability for EFL students.

Oller, Atai and Irvine (1974) re-examined the exact and acceptable scoring procedures for use as measures of EFL proficiency, and found that they intercorrelated at .94. Moreover, although the acceptable procedure correlated higher with dictation (.75) than did the exact word procedure (.69), all other correlations with external criteria - in this case, the TOEFL - were virtually identical. The conclusion reached (contrary to Oller, 1972), was that the exact word scoring procedure measures the same attributes as the acceptable procedure, and is to be preferred on grounds of convenience.

This finding was confirmed by Stubbs and Tucker (1974), who recommended use of the exact word procedure because it correlated at .97 with the acceptable-word procedure. However, as we have already seen in Chapter 2, they overlooked the fact that the acceptable-word procedure had consistently higher validity coefficients. The difference between the correlations of the two procedures with the criterion entrance test of the AUB was approximately of the same order as that found by Oller (1972), which led him to the opposite conclusion, namely, that the acceptable word procedure was the more valid.

In summary, then, for native speakers it seems to be

generally agreed that the exact word scoring procedure is the most appropriate, but with non-native speakers the position is not so clear. Further investigation of this point seems necessary, and, in particular, a comparison of the different procedures with native and non-native speakers, to see if they have a similar effect.

As none of the investigators mentioned attempt to account for the superiority of one or the other scoring procedure in theoretical terms, it will also be necessary to attempt to account for any difference or similarity between procedures that might be found.

3.6 The effect of varying the rate of deletion in a cloze test

Since Taylor (1953) found that deleting every 5th word from text discriminated just as well among subjects as deleting every 10th word, it has been customary to use an every-5th-word deletion rate in cloze studies, for reasons of economy. Nevertheless, there have been many exceptions to this generalisation. Taylor himself, for his 1957 paper, deleted not every 5th word, but every 7th word from text. Although Bormuth consistently deleted every 5th word, Moyle (1970) suggests, without evidence, that a deletion rate more frequent than every 10th word would prove too difficult for young children. Benning (1973) removed every 15th word from her passages, whereas Doyle (1973) deleted every 10th word. Similarly with non-native speakers, Carroll et al (1959), feeling that every 5th word deletion would be too difficult, removed every 10th word instead. Oller (1972) removed every 7th word without justifying his choice of frequency, Anderson (1970) deleted every 8th word, but Stubbs and Tucker (1974) stuck to the more

traditional (for native speakers) deletion of every 5th word. The unspoken assumption in all these cases, however, is that results on one text at one deletion rate are directly comparable with results on other texts at other deletion frequencies; in other words, that the deletion rate used has no effect on the results obtained. However, since both Bormuth (1964c) and Taylor (1954b) found that the five possible different versions of a cloze test at deletion rate 5 produced significantly different results (see section 3.3, this chapter), it might be expected that a different frequency might also result in significantly different results.

A few investigators have, in fact, looked at some aspects of deletion frequency with native speakers. In particular, some have examined the effect of amount of context on the restorability of words. The information theorists were interested in estimates of redundancy of English, which they calculated using the Shannon guessing game (Shannon, 1951), in which subjects guess which letter comes next in a series of letters (and, therefore, words). In particular, Burton and Licklider (1955) attempted to discover the extent to which estimates of the redundancy of English texts are dependent upon the number of preceding letters known to the subject. They gave their subjects varying amounts of preceding context - where Shannon had given 15 and 100 letters of context, they gave subjects the following amounts: 0, 1, 2, 4, 8, 16, 32, 64, 128 and approximately 10,000 letters of context, since they hypothesised that English might be 95% redundant (rather than the 50% redundancy calculated by Shannon) if all possible constraints, including subject matter, style, level of presentation,

etc., were taken into account. However, they found that the constraint imposed by 32 letters was little less than that imposed by 10,000 letters; they concluded that therefore written English does not become more and more redundant as longer and longer sequences of text are taken into account, and suggested that in principle their conclusions also apply when words and even sentences are used as the base units instead of letters. Be that as it may, the main finding is that context of more than 32 letters - i.e., between, say, 4 and 8 words - does not increase the constraint.

Shepard (1963) extended the guessing technique to words, and recorded the number of words that subjects could supply in a given amount of time, to fill a gap with varying amounts of bilateral constraint (words either side of the gap). He found that although subjects were able to find more than one possible word for contexts of up to 40 words (20 words unilaterally), in fact the nature of the curve of words supplied was such that there was only a negligible amount of increase in constraint over spans exceeding 10 words unilaterally - the amounts of context used were 0, 1, 2, 4, 6, 10, 40. These findings would imply that even if there is a difference between cloze tests of deletion rate 12 and deletion rate 14, the difference ought to be minimal, and represent the asymptote of a negatively accelerated curve.

Nicol and Miller (1959) also investigated the redundancy of English, using words as their base units. They took sentences from newspaper articles and deleted words at frequencies of every 5th, 6th, 8th, 10th and 12th word. Subjects were asked to restore the original word (although acceptable synonyms were also allowed). No difference

was found between deleting every 5th word and deleting every 6th word, but differences were found between deletion rates 10 and 12. However, for one text, deletion rate 12 was harder than deletion rate 10. Their design was such that it is impossible to compare deletion rates properly: they compared, for instance, deletion rate 6 on one text with deletion rate 8 on a different text, and then concluded that the deletion rates were different. Moreover, they summed the scores on two different texts to arrive at a mean deletion rate score for deletion rates 5 and 6. The validity of such a procedure is in some doubt. Nevertheless, the authors claim to have established that if a subject has eleven words of context, it is easier for him to supply the 12th word, than it is for him to supply the fifth word after having read only four words.

A somewhat better study was carried out by Aborn, Rubenstein and Sterling (1959), who came to the conclusion that context of less than four words between deletions substantially reduces contextual constraint, and that increasing context between deletions beyond ten words does not increase subjects' abilities to restore the deletion. They used sentences 6, 11 and 25 words in length, and investigated both the effect of position of the deletion within the sentences on predictability (the result was that all positions except the final were equally predictable) and also the effect of bilateral rather than unilateral constraint (here they concluded that bilateral constraint is greater than unilateral constraint). They related their finding of maximum constraint operating with between 5 and 10 words of context to Burton and Licklider's finding that 32 letters represent the maximum amount of contextual constraint. However, their findings are somewhat difficult to relate to cloze because they only used

isolated sentences, rather than continuous text.

Salzinger, Portnoy and Feldman (1962) used a series of passages representing different orders of statistical approximation to English and deleted words from these passages with a frequency of every 5^{th} and every 7^{th} word. There were no significant differences, at any order of approximation to English, between the two deletion rates, and so the authors concluded that "apparently subjects either do not or cannot make use of a context of more than five words on either side of each blank."

Fillenbaum et al (1963) compared deletion frequencies of every 2^{nd}, 3^{rd}, 4^{th}, 5^{th} and 6^{th} words on a transcript of speech. They found a steady increase in both exact word and form class scores at each deletion rate, but interestingly, the largest increase in scores came between deletion rates 2 and 3. Nevertheless, there was also a difference for both scores between deleting every 5^{th} word and deleting every 6^{th} word.

The most complete investigation of the effect of varying deletion rates in cloze tests, however, was made by MacGinitie (1960). He took two passages and subjected them to deletion frequencies of every 3^{rd}, 6^{th}, 12^{th} and 24^{th} word. He then compared only those words which were deleted in all four versions, and discovered a significant difference between deletion rate 3 and the rest, but found no significant differences among the 6^{th}, 12^{th} and 24^{th} word deletion rates. Although only half the contrasts between deletion rate 3 and the rest were significantly different, he concluded that while context of less

than four words between deletions substantially reduces contextual constraint, a distance between blanks of five words or more seems to have little effect on the restorability of the blank. He suggests that the effect of context may be different for different form classes, but nonetheless concludes that the redundancy of English for restorative purposes acts mainly with small segments of speech, and, indeed, that perhaps "the units in which thoughts are composed may seldom be greater than five or six words."

Odom et al (1967) compared deaf and hearing readers on texts with every 3rd, 4th and 5th words respectively deleted, and maintained that there was no general effect of the span between deleted words for either group.

Klare et al (1971) experimented with deletion rates 5, 8, 11 and 14 and found that their subjects attempted every blank, except for the test in which every 5th word had been deleted. They found that their attitude to this deletion frequency was much less favourable than to any of the other deletion rates; they therefore recommend use of a deletion frequency of every 9th or 10th word.

Miller and French (1974) used deletion rates 5, 7 and 10 on science and social science materials, and found that deletion rate 7 was easier than the other two deletion rates for science materials. Although they did not present the statistics, they claimed to find no difference between the three rates on social science materials. However, only one deletion rate - every 10th word - had consistently high correlations with an achievement test criterion. They thus conclude that "an every-10th-word count be used for textual materials that are

fact-laden."

McNinch et al (1974) investigated deletion rates 5, 7 and 9 with science, social science and English materials, and found that varying deletion patterns significantly affects the measurement of readability. They conclude that science materials need a low frequency deletion pattern - e.g., every 9th word - whereas social sciences and English should have every 7th word deleted. Despite their conclusions, they in fact discovered no consistently best deletion patterns for any one subject matter, and although they conclude that different deletion patterns do have an effect on the measurement of readability, it unfortunately does not follow from their findings that the differences between deletion rates that they found are necessarily generalisable to other texts from the subject areas concerned.

Nevertheless, their results are interesting in that they contradict previous research conclusions that the deletion frequency has no effect on cloze scores provided that words are not deleted more frequently than every 5th word, and they encourage speculation not only that different deletion patterns might produce different results but also that there may be an interaction between text difficulty and deletion frequency which may also affect cloze scores.

With non-native speakers of English, only one study of the effect of deletion frequency has been carried out. Haskell (1973) found that his six texts were ranked in the same way whether every 5th word or every 10th word was deleted, and he found no significant differences between passage mean scores for deletion rates 5, 7 or 10. No studies have been made of the effect of changing deletion frequency

on validating correlations of comprehension or language proficiency, nor has any direct comparison been made between the differential performances of native and non-native speakers in this dimension. No attempt has been made to account for those research findings that show no differences among deletion rates and those that do show some differences, or to relate such facts as there are to a theory of what the cloze measures.

3.7 Is cloze sentence-bound?

The results of such research as has been carried out on the effect of changing deletion frequency, in so far as they indicate that, providing at least five words of context are available between gaps, the amount of context does not matter, suggest that cloze tests essentially measure the local redundancy of texts and, more specifically, that cloze scores are not sensitive to contextual clues contained in the more remote context. This suggests that cloze is not capable of testing comprehension of a whole passage, at least when that comprehension is dependent on the interrelationships of ideas and sentences. Similarly, the apparent lack of difference between scoring procedures which allow only the exact word, and those which also allow any contextually acceptable word, suggests that if both procedures are measuring essentially the same thing, and the any-acceptable-word procedure basically ignores long-range contextual constraint, or at least, if one can claim that the exact word procedure would normally be expected to measure greater sensitivity to remote constraint, as well as factors like style, register, background knowledge and so on, then

perhaps the cloze really is not sensitive to such considerations. These two apparent findings - which, of course, need experimental verification - suggest the hypothesis that cloze is essentially sentence-bound. The findings reported in Chapter 1, section 1.7 (Reading Gain), that although the rational cloze may be a useful measure of information gained from reading a text, the random cloze does not seem to be suitable for such a purpose, lend themselves to an interpretation that cloze is essentially a measure of the immediate constraints in text, and not of overall comprehension. Thus subjects who read unmutilated text, and then do a cloze test have in fact gained information from the text, and may well have "understood" it, but the random cloze is simply insensitive to such events, and therefore an unsuitable measure of them.

J. B. Carroll, who has always been sceptical about the value of the cloze procedure as a measure of reading comprehension or foreign language ability, claiming that cloze probably involves an ability specific to the procedure itself, rather than one closely related to other verbal abilities (Carroll et al, 1959), suggests, in Carroll and Freedle (1972), that cloze scores are largely dependent on "local redundancy", which he explains as meaning "the extent to which linguistic clues in the immediate environment of a missing word tend to supply it." If this is true, then cloze can hardly be taken as a measure of general reading comprehension, since there is much more to the understanding of a text than the understanding of the immediate environment of words. It is important, for instance, to make connections between sentences even if those relationships are not made explicit by the writer. It is important to relate ideas in one part of the text to ideas expressed in

another part. Equally, it is important to be able to evaluate the relative importance of the different ideas/topics mentioned in text, in order to gain an overall and balanced view of what the author is saying. Clearly there is far more to reading than the comprehension of isolated sentences. However, Carroll concludes that "there is no clear evidence that cloze scores can measure the ability to comprehend or learn the major idea or concepts that run through a discourse."

Several experimental studies throw some light on the question of whether cloze can measure this ability, and in particular, they help to inform any discussion as to whether cloze is largely a measure of local redundancy - i.e., constrained by the immediate environment - or whether it is sensitive to constraints from remoter parts of the text. If the latter should prove to be the case, then it is more likely that cloze is, at least in principle, capable of measuring "overall comprehension". If the former is the case, viz., that cloze is a measure of local redundancy, then it is probably in principle incapable of measuring "the ability to comprehend or learn the major idea or concepts that run through discourse".

Coleman and Miller (1967) used three variations of the cloze procedure in order to estimate the readability of 36 passages. Cloze procedure type 1 was the ordinary procedure, deleting every 5^{th} word from the texts. This deletion system was carried out five times on each text; once deleting the first word, and then every fifth word thereafter; once deleting the second word, and then every fifth word; once deleting the third word and then every fifth word; the fourth word and every fifth; and the fifth word and every fifth word. In this way

a cloze score was obtained for every word in the passage.

The second deletion system, type 2, involved the deletion of one and only one word from the text. The texts were 150 words long, so 150 different versions of this procedure had to be prepared to gain a type 2 cloze score for every word.

The third deletion system, type 3, was intended to restrict the context available to what precedes the item. In this procedure, the subject is given the first word, and told to guess the next word. After his guess, correct or incorrect, the correct word is revealed, and the subject must then try to guess the next word. This he does until he has attempted every word in the text.

The three cloze scores for each word were averaged for each word, and then for each text. It was also possible to examine the difference between words depending on the position the word held in a sentence.

Taking sentences of eight or more words, Coleman and Miller compared the average cloze score of the first, second, third and fourth words in a sentence, and the last, next-to-last, third-from-the-end, and fourth-from-the-end words in the same sentence, and found that there was a steady increase in the cloze score as the position of the cloze item neared the end of the sentence. They concluded that there is a high sequential constraint on words within sentences.

To investigate the constraint on words operating from outside the sentence, they compared the average cloze scores from the twentieth word in a text (i.e. almost certainly in the second sentence) to the last word in the text. If there is constraint beyond the sentence,

then the cloze scores should increase over this range. As no such increase was found - socres being approximately steady from 20th to 140th position - the authors conclude that "very little constraint comes from words in other sentences". The general conclusion is that cloze scores are very largely influenced by within-sentence constraints, and hardly by between-sentence constraints.

Interestingly, the mean word scores for the type 1 deletion were 54.6%, whilst the mean score for type 2 deletion was 63.8%. In other words, there is not a very great increase in ease if the subject has the remaining 149 words of text rather than the text left after every fifth word has been deleted. This, again, adds weight to the claim that cloze is not greatly influenced by relatively remote context.

Several studies looked at the effect of context on cloze scores. Musgrave (1963) tested the supposition that text may be "correctly understood more often" if presented in context, by giving subjects a cloze test and presenting three of the four groups with an unmutilated lead paragraph. In one case, this paragraph contained information about the person in the mutilated text and the subject matter (the "who" and "what"). Another group received a lead paragraph containing only "who" information, and a third group received only "what" information. No significant difference was observed between any of the four conditions, and Musgrave concluded that, since other testing techniques are able to show the effect of adding "who" and "what" information to texts, the cloze technique is simply not sensitive enough to such additional context, possibly, she suggests, because cloze does not take into account "the kind of meaning and cognitive content

carried by 'who' and 'what' context."

Erickson and Hansen (1974) took two texts, selected five 60 - 70 word passages from each text, and deleted every 5th word from these selections. In one condition they simply presented these passages, in the order in which they had occurred, to subjects. In the other condition they surrounded each paragraph by the approximately 75 words of original text that had originally enclosed it. This, the "in context condition", thus gave 50 cloze items in the form of 10 deletions followed by an intact section followed by 10 deletions, and so on. They found no significant difference between the two conditions - in fact, there was a non-significant tendency for the "out of context" condition to be easier. They thus concluded that cloze was not sensitive to non-immediate context. They also investigated a suggestion by Ramanauskas (1971) that cloze responses at the beginning and end of a cloze test should be compared to see whether there is a cumulative effect of context clues - if context has an effect, the text should become more predictable, and thus cloze item facility should rise towards the end. They were unable to demonstrate such an effect, however, since there was no significant difference between the mean score for the first 10-item passage of their tests and the last 10-item passage. It is, of course, possible to conclude, not that contextual constraint does not increase as one goes through text, but that cloze is not sensitive to such increasing constraint.

Suhorsky (1975) took a passage which he subjected to deletions of every fifth word. He prepared three different versions of his test by adding to the beginning of the test, in the first case six T-units

(a T-unit is a main clause, and any related subordinate clauses), in the second case one T-unit, and in the third case, nothing. This corresponds roughly to 1) an ordinary cloze test of 50 items 2) the same test with one sentence of preceding context, and 3) the same text with a paragraph of preceding context. When he compared the three tests, he failed to find significant differences, and thus concluded that "isolating a text from context" has no effect.

A somewhat more ambitious study was undertaken by Bartoo (1975). Three of his cloze test were as follows:

1) 300 words from the end of a passage, subject to a deletion of every fifth word;
2) as in type 1 but preceded by the immediately previous 300 words from the passage, unmutilated; and
3) as in type 1, but preceded by the immediately previous 600 words.

Having administered these tests to tenth-grade students, he discovered that there was no significant difference among conditions, and so concluded that adding even six hundred words of relevant context to a cloze passage had no effect on the cloze scores. However, if the cloze procedure is a measure of reading comprehension, one would expect scores on a passage to be influenced by the amount of information relevant to the passage that was available. That this is not the case appears to indicate that cloze is not greatly influenced by context beyond the immediate sentence.

To investigate Carroll's contention that cloze scores are largely dependent on local redundancy, Tuinman, Blanton and Gray (1975) took a text, which they called UA, and reduced its "structural redundancy" by varying amounts. In the Text M1, 30% of the text was

removed by deleting all function words defined in a narrow sense (Coleman 1971). Text M2 was a 50% reduction, obtained by deleting all function words defined in a broad sense (Coleman 1971). These three texts, UA, M1 and M2 were given to seventh-grade children, either accompanied by 32 multiple-choice questions, or subjected to an every-5th-word deletion. The reduction of redundancy as represented by Texts M1 and M2 had relatively little effect on comprehension as measured by the multiple-choice questions (Text M1 had a mean which was 87% of UA, Text M2 had a mean which was 77% of UA), whereas the cloze scores were drastically lower than the scores on the original text (M1 mean was 22% of UA, M2 mean was 25% of UA). The authors conclude, not only that the reduction of redundancy has a great effect on cloze scores, but also that, since the cloze scores were very low, the lack of structural redundancy means that subjects cannot cope with cloze. Their general conclusion is that "performance on cloze tests on intact passages depends to a considerable extent on a subject's ability to utilize . . . structural redundancy, rather than on a conceptualization of the message content of the passage, such as is the case for comprehension measures utilizing questions."

Carroll et al (1959) used native speakers of English to test whether context clues from the whole passage affect restoration on a cloze test. They took 20 ten-word fragments from three texts, and deleted the fifth word from each fragment. These fragments they arranged randomly, and compared subjects' ability to restore these deletions with performance on the items in their original order in text. They found a significant difference between the two conditions,

when the ability to do cloze was held constant. They did not, however, find a cumulative effect of clues, since the mean of the first eight items was no lower than the mean of the last eight items, both on scrambled and unscrambled text.

Moreover, they found that only 26 out of the 60 items showed a significant difference between scrambled and unscrambled text, and three of these were in the wrong direction (easier when scrambled). The relative difficulty of items did not change on scrambling, and there was no absolute change in the difficulty of deleted prepositions and adverbs. Only some (6 out of 13) nouns and some (6/8) verbs were harder to restore when in a scrambled text. They conclude that paragraph clues do affect cloze scores, and that cloze is therefore sensitive to more than the immediate environment. However, their study can be criticized because the scrambled tests used fragments of language - four words, a gap, then five words - regardless of syntactic boundaries. In view of the considerable evidence that text processing occurs in syntactic units, probably the clause (see Foder, Bever and Garrett, 1974), it is possible that this procedure disturbed the processing process, and that if, for example, complete sentences had been used, the results might have been different. In any case, their results do not negate the thesis that cloze is sentence-bound. In fact, their evidence shows that most cloze items (34 out of 60) are not affected by more than five words of immediate context.

In order to look at the effect of context beyond the sentence in which the cloze item is to be found, Ramanauskas (1971, 1972) applied the cloze procedure to a text (deletion rate every fifth word),

and then took the mutilated text, and randomly rearranged the sentences, retaining the same deletions, of course. She found significant differences between her conditions, and concluded that constraints beyond the sentence do operate on cloze items.

However, there are two drawbacks to her study. The first is that her subjects were educationally subnormal children with a reading grade of 2.5, which makes it extremely difficult to generalise her claims to normal native speaker readers, not to mention non-native speaker readers. The second drawback is that she used two texts, and in her modified tests (MOD), with randomly rearranged sentences, she intermingled half the rearranged sentences from one text with half the rearranged sentences from the other text to create her test. The sentences were not presented as separate sentences, divorced from context, but were placed together, in one paragraph, as if forming one coherent text. This might well have had confounding effects.

The other study using randomly rearranged sentences was by Marshall (1970), using as subjects deaf and hearing children with a normal IQ. His tests involved three levels of contextual constraint, which he called 1) discourse level (i.e., the original text); 2) discrete sentence level (i.e., the text with its sentences randomly rearranged); and 3) fragment level, where only a minimal number of contextual clues were retained "so as not to destroy the earmarks of any given grammatical construction" (this can perhaps be thought of as "phrase level"). Although his cloze tests based on these differing levels of contextual constraint gave him overall differences in conditions, a closer inspection revealed that the only differences were

between "the higher scores of the connected discourse level, and the lower scores at the fragment level". Therefore, no difference was found between levels 1 (discourse) and 2 (discrete sentences). It seems also fair to conclude that although the cloze scores may not have been entirely determined by the minimal context of the fragment level, it does not follow that the determinant is beyond the sentence. It is important to note that these results were true, not only for the cloze form class score - i.e., a noun is identifiable as a noun within the sentence - but also for the cloze exact word score. Thus, on the question of the range of cloze constraints, the evidence as to whether they range beyond the sentence is contradictory, and the area would benefit from further investigation.

Finally, Ferry (1975) used the cloze procedure to investigate the readability of text with and without coherence markers (words signalling "relationships between sentences"). He took a passage which he assumed to have medium coherence marker density, and created two other texts: a high density coherence marker text, by adding coherence markers, and a low density coherence marker text, by deleting coherence markers. The three texts were subjected to deletion of every fifth word, giving 80 items per test, and the tests given to tenth-grade students. No significant difference between texts was found. It is therefore possible to conclude that cloze is not sensitive to enriched or impoverished coherence, and from this, to doubt whether cloze is at all sensitive to coherence. If it is not sensitive to coherence, it is unlikely to be a suitable measure of reading abilities.

To summarise, then, studies have shown that cloze tends to

measure structural redundancy and its utilization, rather than comprehension of the discourse. They have shown that cloze is insensitive to discourse enrichment, in terms of coherence markers, and it is insensitive to increases in (relevant) context. Despite the doubt which still exists as to whether constraint operates on cloze beyond the sentence as measured by the rearrangement of sentences, there is strong evidence that the major constraints operating on a cloze item are within the sentence, rather than outside or beyond the sentence.

The evidence tends to support Carroll's claim that cloze depends largely on local redundancy. Whether that redundancy operates solely within the confines of the sentence, or whether it is simply a question of amount of context regardless of sentence boundaries, remains to be investigated. Of course, these results are not entirely surprising, since the cloze procedure, as currently used, involves the deletion of words, rather than phrases or ideas. One would probably expect that beyond-sentence constraints could only operate on cloze in so far as they can operate on individual words, or in so far as discourse coherence can be carried by individual words such as coherence markers, anaphoric items, or identical lexical items. It is difficult to imagine the deletion of a word which would tax one's ability to make inferences, which would require an inference on the part of the reader to restore, or which would measure other higher-order comprehension skills like evaluation or identification of the main idea.

No studies of the sentence-bound nature of cloze are known to this author using non-native speakers of English as subjects.

If, as was suggested at the beginning of this section, it is

found that there are no significant differences between cloze deletion rates, this can be taken as evidence that cloze is essentially a measure of local redundancy, of the constraints of the immediate environment, possibly at a level lower than the sentence. This study hopes to throw some light on the nature of what cloze tests by examining different deletion frequencies, although it will not investigate other aspects of the same problem by, for example, comparing scrambled and unscrambled cloze items. It is hoped that the study will also indicate whether what is true for native speakers also holds for non-native speakers.

CHAPTER 4

Pilot Study — Algeria

4.1 Aim

In view of the previous research reviewed in Chapters 1 - 3, it was felt that attempts to treat the cloze procedure as unitary were misleading, and that research was needed, especially with non-native speakers, into the effects of varying some of the variables seen to be of importance in Chapter 3, to see if different versions of the cloze procedure were comparable.

It was decided to carry out a pilot study of the effect of changing the cloze deletion rate on the measurement of text readability and the measurement of proficiency in English as a foreign language, with non-native speaker subjects. If results proved interesting, then a closer, in-depth study would be undertaken to follow up such avenues of research as might prove to be of value. Since this was a frankly exploratory study, no formal hypotheses were set up, but there were several expectations: 1) that changing the deletion rate would affect the measurement of the difficulty, for non-native speakers, of reading passages; 2) that varying the deletion rate would also affect the estimates of proficiency in English as a foreign language for any non-native speaker; 3) that texts of differing difficulty would measure English proficiency differently; and 4) that readability formulae are of little value in determining the difficulty of text for non-native speakers of English.

4.2 Subjects

The subjects of this study were students of English studying in the English Department of the University of Algiers, where the author was a lecturer. The students, some 75% of whom were female, were aged between 18 and 50, with the majority in their late teens and early twenties. They were native speakers of one of three languages: Algerian Arabic, French, or Kabyle (a Berber language). For all the students, either French or Arabic was their second language, and virtually all were at least bilinguals (many Kabyle speakers were effectively trilingual.) English was being learned as a foreign language in the English Department, and most of the students intended to teach English at secondary or tertiary level thereafter. Many of the students had studied English in school for up to six years, but there were also several virtual beginners in the sample. The University course lasted for three years, and the tests were administered to each of the three year-groups.

4.3 Materials

Twelve texts were selected to give a range of apparent difficulty as measured by the Fog index of readability (Gunning, 1952), from easy to very difficult. The texts were chosen from various sources - newspapers, literary essays, academic essays, prose fiction - and corresponded in general to what the English students of Algiers University were expected to read during their courses, both in subject matter and difficulty. (For the intact texts, see Appendix A.)

Each text was approximately 300 words in length, partly in order to accommodate it onto one sheet of paper in the test booklet,

and also in order not to have an overwhelming number of deletions at the more frequent deletion frequencies.

Each text was paired off with another of approximately equal difficulty, and given a letter from A to F. To distinguish the two in the pair, the subscript 1 or 2 was given.

Every n^{th} word was then systemically deleted from the texts, at six different deletion rates - every 14^{th}, 12^{th}, 10^{th}, 8^{th}, 6^{th}, and 4^{th} word - giving 72 tests in all.

At any given level of difficulty there were therefore twelve tests: A1 (14), A1 (12), A1 (10), A1 (8), A1 (6), A1 (4),
A2 (14), A2 (12), A2 (10), A2 (8), A2 (6), A2 (4).

Each student would be given a booklet containing six tests, covering all six levels of difficulty, and with one example of each deletion rate.

To control for order of difficulty affecting the performance of the students the order of the six levels was permutated in the following manner:

ABCDEF, ADBECF, ACEDBF, AFDCBE, ADECFB, etc.

Next, to ensure that text A1 was not always contained in the same booklet as B1 or F1 (etc.) the subscripts 1 and 2 were ordered in all possible combinations, e.g.,: 111111 (and its mirror image 222222), 121212 (212121), 111222 (222111), 122211 (211122), etc.

Then, to avoid text A always appearing at the 14^{th} rate of deletion and text F at the 4^{th} in the same booklet, the deletion rates were controlled through rotation, thus:

Booklet 1: A1 (14), D2 (12), B1 (10), E2 (8), C1 (6), F2 (4)

Booklet 2: A1 (12), D2 (10), B1 (8), E2 (6), C1 (4), F2 (14)

Booklet 3: A1 (10), D2 (8), B1 (6), E2 (4), C1 (14), F2 (12)

Booklet 4: A1 (8), D2 (6), B1 (4), E2 (14), C1 (12), F2 (10)

Booklet 5: A1 (6), D2 (4), B1 (14), E2 (12), C1 (10), F2 (8)

Booklet 6: A1 (4), D2 (14), B1 (12), E2 (10), C1 (8), F2 (6) etc.

Because, during the test, it might have been possible for student 1 to receive A1 (14), and his neighbour A1 (4), thereby allowing student 2 to copy from student 1 even if the latter wrote nothing, the booklets were arranged and distributed, as far as possible, in mirror image pairs. So student 1 might receive the following booklet:

A1 (4), D2 (14), B1 (12), E2 (10), C1 (8), F2 (6)

and his neighbour the following:

A2 (4), D1 (14), B2 (12), E1 (10), C2 (8), F1 (6).

Finally, the test booklets having been compiled to ensure that, for example, deletion rate 14 was not always followed by deletion rate 12, a random selection of possible permutations was made (14, 12, 10, 8, 6, 4; 4, 6, 8, 10, 12, 14; 8, 4, 12, 6, 14, 10; etc.), and these permutations were then rotated in order (14, 12, 10, 8, 6, 4; 4, 14, 12, 10, 8, 6; 6, 4, 14, 12, 10, 8; 8, 6, 4, 14, 12, 10; etc.) to control for order effect of the rate of deletion.

The result was that no two booklets were the same.

In order to compare the cloze tests with a known and trusted measure of proficiency in English as a foreign language, the English Proficiency Test Battery (EPTB, or Davies Test), Short Version, Form B, 1965, Part One, was also administered to the students. The battery is used by the British Council abroad for screening foreign students

intending to study in Britain, and it is possible to determine whether students are too weak in English to benefit from study, whether after a short remedial course their English would be adequate, or whether they have a level in English which suggests that they are not at risk linguistically. Part One of this battery consists of 4 subtests. Test 1 is a phoneme discrimination test, in which students have to say whether the words in the triplets they hear are the same or different. Test 2 is said to be an intonation test, but it approaches a test of general listening comprehension, in which students have to interpret a dialogue correctly, usually involving understanding of some crucial stress or intonation pattern. Test 3 is a modified rational cloze test - only function words are deleted from two short passages and the initial letter of each deleted word is retained as a guide for students. Test 4 is a traditional multiple-choice test of selected grammatical points which are felt desirable for foreign students to master.

4.4 Administration

When giving its permission to carry out the investigation, the English Department insisted on two things: 1) that it be made clear to students that participation was entirely voluntary, and 2) that no special accommodation or timetabling could be contemplated. Even if the second condition had not been insisted upon, the existence of the first would have resulted in a bias of (presumably good) volunteers, or resulted in no students at all participating. This led to two problems. Firstly, it was necessary to hold the tests during normal class hours, which meant that the only people tested were the students who attended class on that day. (No warning was given, or students might have stayed

away from class.) There was therefore greater participation on the part of first-year students than on the part of second- or third-year students. Secondly, because students could not be compelled to take the tests, many who started became bored and left before the two hours allotted had expired. This resulted in some tests being incomplete and some being left blank (although all students claimed they had done their best, and it may perhaps be assumed that those tests left blank were felt to be especially difficult).

The first series of cloze tests were given to all three groups the week of May 20 - 24, 1974, during the Compréhension Ecrite class (two hours), under supervision, usually, but not always, by the author. The English Proficiency Test Battery (EPTB), was given separately, during the same week, to all students who attended the language laboratory class of Compréhension Orale (one hour). This was administered in the language laboratory by the author.

Later, an attempt was made to induce those who had taken cloze but not EPTB to take EPTB, and those who had taken EPTB but not cloze to take cloze, during the week of June 17 - 21, but because by now students were forewarned and the operation was clearly voluntary, this resulted in only 36 extra cloze testees, and 40 extra EPTB testees.

4.5 Results - General

Ignoring anonymous booklets, joint efforts and duplicated copies (where students took the cloze test twice), 242 students took the cloze tests, and 243 took the EPTB. Of the latter, some 104 did not take the cloze, and so were discounted for the study. Thus, 139 students took both cloze and EPTB, and 103 took cloze only. These 242

students were divided amongst the three year-groups as follows:

First Year	Second Year	Third Year	Unknown	Total
76	78	85	3	242

Between 14 and 26 copies of each test were retained (the number varied because of the way the booklets were made and distributed, but the average was 20). The details are given in Table 4.1, where it will be seen that approximately 120 copies of each text were obtained and, since every subject took one copy of each deletion rate, 242 copies of each deletion rate were completed.

T A B L E 4 . 1

Number of students taking each cloze test

Text DR	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	F2	Total
4	17	19	23	14	21	20	21	19	23	21	21	23	242
6	21	20	21	21	19	20	20	23	20	18	21	18	242
8	21	24	19	23	19	16	19	22	19	21	20	19	242
10	18	17	19	19	26	19	21	16	25	23	22	17	242
12	21	19	19	24	20	20	20	20	17	18	22	22	242
14	25	20	21	19	21	21	23	18	18	19	18	19	242
Total	123	119	122	120	126	116	124	118	122	120	124	118	1452

Table 4.2 below presents brief details of each text, and its Fog readability index.

TABLE 4.2

The nature of the texts

A1:	prose fiction - short story - Shadbolt	Fog: 7
A2:	newspaper column - humorous essay - J. Cooper	Fog: 8.7
B1:	prose fiction - short story - Rhys Davies	Fog: 8
B2:	essay - socio-political (literary) - Orwell	Fog: 11
C1:	essay - socio-political (literary) - Orwell	Fog: 14.2
C2:	newspaper report - political speech - Sunday Times	Fog: 13.9
D1:	newspaper editorial - political comment - Guardian	Fog: 13.5
D2:	prose fiction - short story - Rhys Davies	Fog: 14.3
E1:	newspaper article - educational/psychological - Sunday Times	Fog: 17.5
E2:	prose fiction - short story - Gordimer	Fog: 16
F1:	essay - socio-political (academic) - Krausz	Fog: 21.5
F2:	newspaper essay - television critic - Sunday Times	Fog: 17.3

In view of the large number (72) of different tests, the cloze tests were only scored by the exact word method - i.e., only responses which exactly restored the word deleted (with minor spelling errors ignored) were judged to be correct. Because the same length of text was used for each deletion system, different deletion rates resulted in a different number of deletions per test. To enable different test scores to be compared, therefore, all raw scores were expressed as a percentage of the total number of deletions in any given test, and these results are set out in Table 4.3, which gives the mean cloze exact score for each text, at each deletion rate.

T A B L E 4 . 3

Text means by deletion rates:

	14	12	10	8	6	4	Total
A1	28.0	25.2	26.6	25.1	22.4	16.8	24.0
A2	41.5	18.7	35.6	28.7	26.9	27.1	29.7
B1	22.6	18.9	19.5	15.6	14.6	10.0	16.9
B2	31.1	41.3	26.1	36.0	30.2	22.2	31.2
C1	24.8	20.0	24.1	27.2	23.8	17.5	22.9
C2	19.0	38.2	33.2	28.0	26.0	25.2	28.3
D1	27.9	20.2	20.3	20.6	13.2	18.6	20.1
D2	29.4	21.9	17.0	19.7	19.6	14.3	20.3
E1	20.7	20.4	27.8	20.2	17.5	11.9	19.8
E2	20.8	22.0	27.7	30.8	29.9	17.7	24.8
F1	23.7	24.7	31.8	20.4	20.9	19.2	23.5
F2	24.0	27.5	19.5	19.5	17.2	12.6	20.1
Total	26.13	24.92	25.77	24.32	21.85	17.76	23.46

A clear range of scores was achieved, from 41.5% mean correct restorations on Text A2 with every 14^{th} word deleted, to a mere 10% mean correct restorations on Text B1 with every 4^{th} word deleted.

The results of those 139 students who also took the EPTB tests are presented in Table 4.4. It should be noted that the standard score mean and the standard deviation for the whole test, reported by its designers, are 40.00 and 6.00 respectively. Thus the mean of 40 achieved by the Algerian students suggests that their average ability is typical of the population of foreign learners of English who take EPTB.

A somewhat lower standard deviation - 4.12 - suggests that this sample of students is less heterogeneous than the population. This would appear to be reasonable, in view of the fact that they all study at the same university, and are from similar linguistic, educational and cultural backgrounds.

TABLE 4.4

Performance of 139 students on EPTB, Form B, Part One

n = 139

	Mean	Standard deviation	Range	Maximum possible
Total	40.02	4.12	26 - 51.6	55.7
Test 1	9.96	1.55	3.5 - 12.4	13.1
Test 2	10.07	1.40	6.6 - 13.5	15.1
Test 3	10.10	1.57	6.3 - 13.6	14.4
Test 4	9.90	1.11	7.1 - 12.7	13.1

The somewhat lower than normal spread is reflected in the standard deviations for the individual tests. The standard standard deviation is reported as 2.00, with a mean of 10.00. These subjects' standard deviations vary from 1.57 to 1.11. Interestingly, whereas performance on the two listening tests (Tests 1 and 2) was almost exactly normal, performance on the modified cloze (Test 3) was marginally above average whilst performance on the grammar test (Test 4) was marginally below average.

In general, however, it appears that the subjects show no abnormalities in their English language abilities, that they can be taken

to be representative of those learners of intermediate to advanced level who take the EPTB test, but that they are more homogeneous than the population. In fact, a standard score of 40 (the mean achieved by the subjects) is interpreted as indicating that a student has sufficient English to enable him to study in the United Kingdom without the need for remedial tuition of any nature.

The following two sections examine the cloze results to determine the influence of the experimental variables of deletion rate and text.

4.6 The effect of changing the deletion rate

Several analyses were carried out on the data in order to determine the effect of varying the deletion rate.

1) Subjects having taken one test at each deletion rate, the scores were ranked for each individual, and the total rank values for each deletion rate were calculated. The Friedman Two-Way Analysis of Variance, and the Selected Pairs Comparison Test (Langley, 1970) were applied to test for significance of the difference between ranks for each treatment. The different texts were assumed to be equivalent. (Tables 4.5 and 4.6)

The subjects were divided into groups according to different criteria (by year, by score on EPTB, and by whether or not they had taken EPTB) and the rank totals for deletion rates were compared (Tables 4.5 and 4.6).

2) A t-test for significance of the difference between

means of deletion rate scores (for correlated samples) was applied to the data for all subjects. The different texts were assumed to be equivalent (Table 4.7).

3) Graphs were plotted from the mean scores on deletion rates for all subjects (Figure 4.1), for scores on deletion rates on easy texts compared with difficult texts compared with intermediate texts (Figure 4.2), for the easiest six texts compared with the most difficult six texts (Figure 4.3), and for each text individually (Figure 4.4, a and b).

4) t-tests for independent samples were applied to the data by text in order to test for significant differences between means of deletion rates on each text. (Table 4.9)

4.6.1 Results

Since each subject had taken one example of each deletion rate (deletion rate = 14, 12, 10, 8, 6, 4), 242 sets of matched measurements had been obtained. Friedman's Test was used to test for significant differences among deletion rates - i.e., would tests with less frequent deletion of words prove significantly easier to complete than tests with a more frequent deletion? (One, would, of course, expect that the less contextual constraint surrounding each blank, the more difficult it would become to restore the deletion.)

Highly significant differences ($p < .001$) were found among deletion rates for all subjects (Table 4.5a), and for 1^{st}, 2^{nd} and 3^{rd} years individually (Table 4.5 b, c and d). Similarly, the difference between rank totals was highly significant for those students who had taken EPTB, and for those who had not (Table 4.5 e and f). However,

those students who scored highly on EPTB showed a less significant difference among their cloze deletion rates ($p < .05$), whilst no significant difference between their cloze scores was found for those who scored badly (one standard deviation or more below the mean) on EPTB (Table 4.5 g and h). This latter result is somewhat surprising, since one would expect weak students to need more contextual constraint than average students, and thus to perform significantly better on less-frequently-deleted texts. It is possible that these tests were simply too difficult for the weak students, even at deletion rate 14. Conversely, one would expect the better students to need less textual information, and therefore their performance on varying deletion rates to vary less than that of average students. This proved to be the case. (However, caution must be applied when interpreting the results of Table 4.5 g and h, since the number of students is low, owing to the relative homogeneity of the students regarding their English proficiency as measured by EPTB.)

When one looks at the relative ranking of the total rank values, one sees that the only consistent pattern is that the deletion of every fourth word is always more difficult than the deletion of fewer words, whether one looks at the data for all students, or for any of the sub-groups mentioned above. Although there is a general tendency for the difficulty of the test to increase as the frequency of deletions increases, it is by no means consistently true that a deletion rate of 14 is easier than 12 is easier than 10 is easier than 8 is easier than 6; it is especially difficult to discern a pattern when the number of subjects is low. On the assumption that the most reliable data is that

which includes most information, one should take Table 4.5a (all students) to be the best guide to the effect of deletion variations on comprehension. And here it appears that it is easier to complete cloze tests where every 10^{th} word has been deleted than ones where every 14^{th} or every 12^{th} word has been removed.

Where the Friedman analysis revealed significant differences overall, a further analysis was undertaken to attempt to discover the source of this significant difference. This was done by taking pairs of scores and testing for significant difference (Table 4.6). From this tabulation it is immediately apparent that the difference mentioned above between deletion rate 10 and deletion rates 12 and 14 is, in fact, not significant, and so could be attributed to the vagaries of chance. More striking is the fact that, with non-patterned exceptions, the only significant differences to emerge are between the fourth deletion rate and all the rest. There is no significant difference between deletion rates of 14, 12, 10 and 8. Thus the <u>amount</u> of contextual constraint exerted on a word would appear from this first analysis to be irrelevant provided that it is not less than five words between blanks. This appears to confirm previous findings by MacGinitie with native speakers, and Haskell with ESL students.

2) The findings from the initial analysis are not entirely supported by a t-test for correlated samples applied to the deletion rate data for all subjects. Whereas the Selected Pairs Comparison Test had found virtually no significant differences between deletion rate 6 and less frequent rates, the t-test reveals highly significant differences ($p<.01$, Table 4.7).

T A B L E 4 . 7

t-test for correlated samples on deletion rates, all subjects

Deletion rates	14	12	10	8	6	4
14		NS	NS	NS	++	++
12			NS	NS	++	++
10				NS	++	++
8					++	++
6						++

NS Not significant at 5% level ($p > .05$)

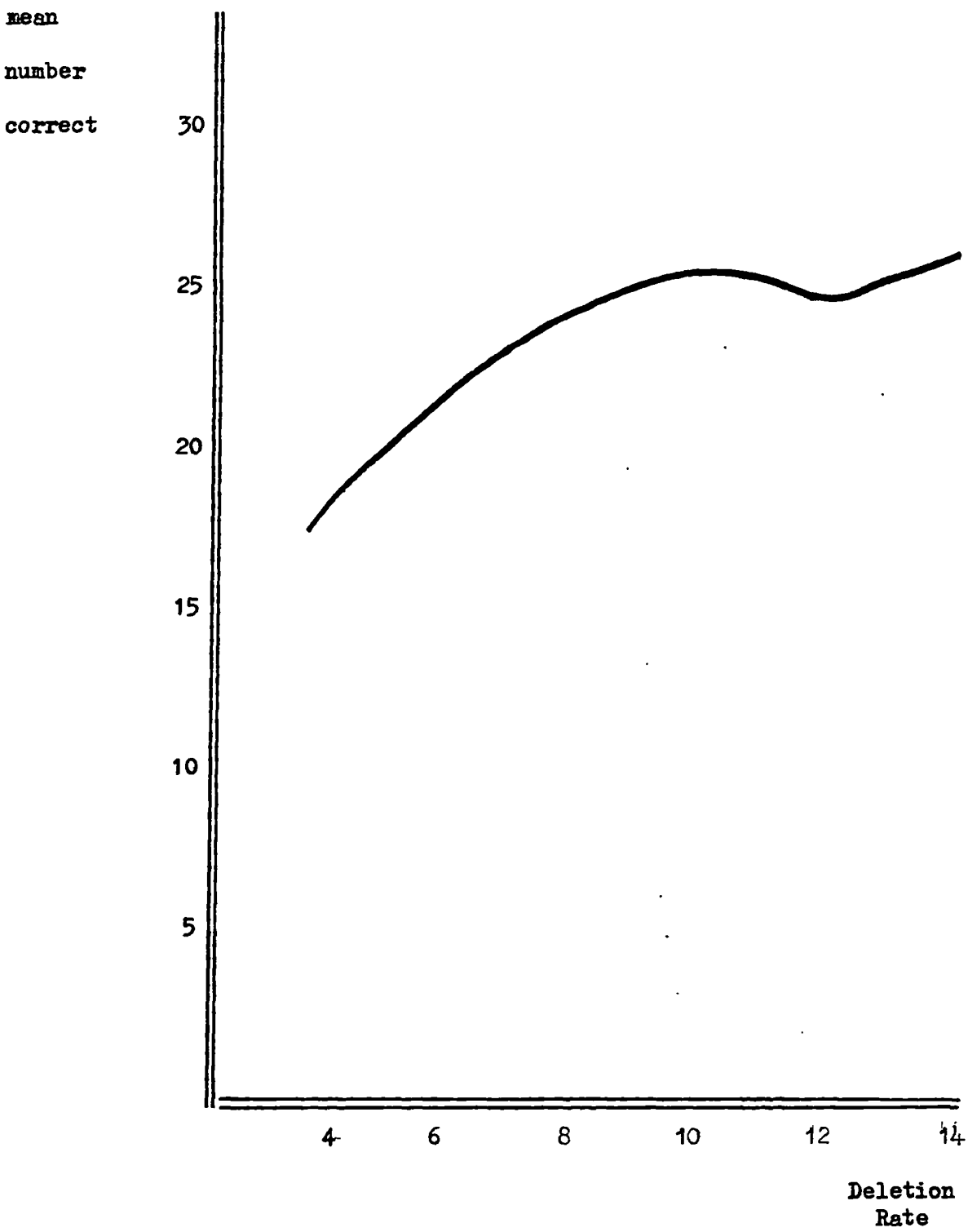
++ Significant at 1% level ($p < .01$)

This suggests that one should have at least seven words of context between blanks, and not five as suggested by the initial analysis. This, of course, is not as easy to explain in terms of MacGinitie's or Haskell's findings. Nevertheless, it still appears to be true from the results of this t-test that it does not matter whether one deletes every 8^{th}, 10^{th}, 12^{th} or 14^{th} word.

3) If one plots the cloze scores onto a graph as a function of the change in deletion rate, one can get an idea of the <u>tendency</u> of comprehension to increase or decrease with varying amounts of contextual constraint, regardless of <u>significant</u> differences. If there is no difference between deletion rates, one would expect a more or less horizontal line; if there is a constant increase in difference amongst rates, one would expect to find some kind of diagonal line. Figure 4.1 shows that the increase in percentage of text restored, regardless of text, is a negatively accelerated curve which levels out at deletion rate 10.

F I G U R E 4 . 1

Overall cloze score, exact word procedure, regardless of text (Algeria)

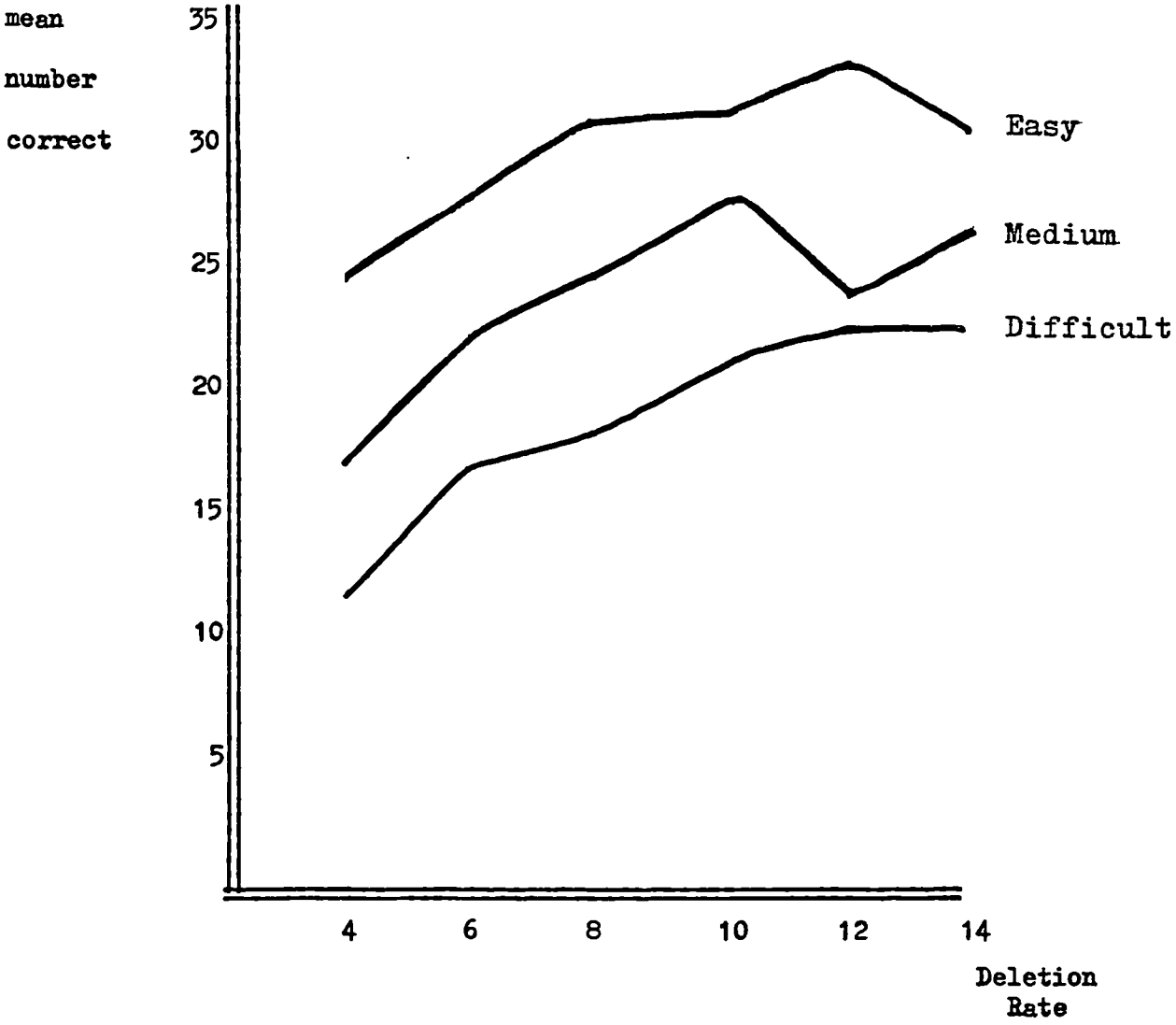


It would appear from this that although differences between deletion rates may not be significant, there is a distinct tendency 1) for scores to increase up to deletion rate 10, and 2) for there to be no difference between deletion rates 10, 12, and 14. In other words, perhaps there should be at least nine words of context between blanks for the amount of contextual constraint to become irrelevant.

However, like Tables 4.5 to 4.7, Figure 4.1 is based on the assumption that there is no important effect contributed by the difference in texts. Figure 4.2 attempts to take account of some of the difference in texts by grouping texts according to difficulty as measured by the cloze score, not the Fog Index, and showing the increase in cloze scores for the easiest three texts against those of the most difficult three texts, and those of the three intermediate texts.

FIGURE 4.2

Cloze scores (easy texts vs. difficult texts vs. intermediate texts) by deletion rate (Algeria)

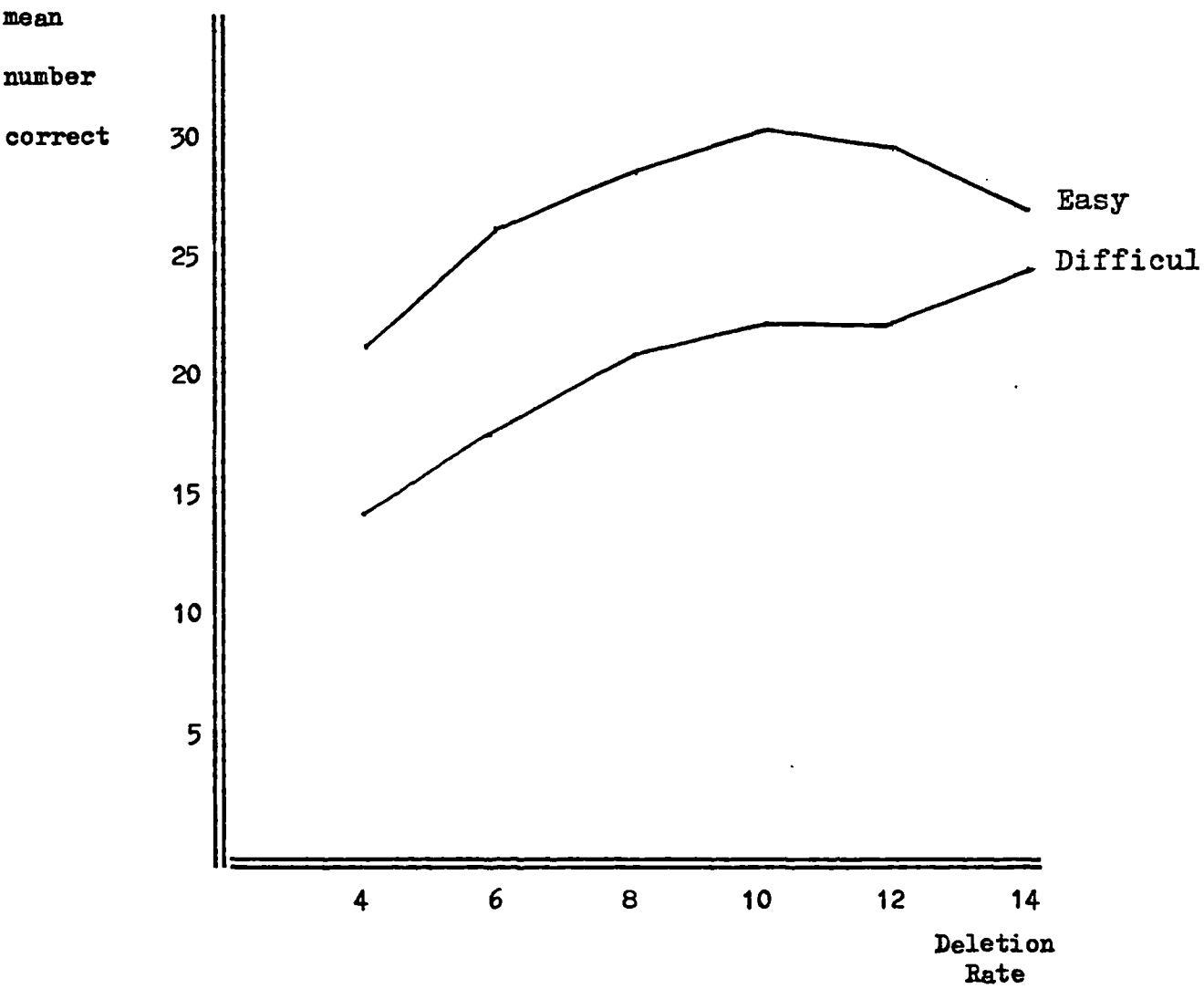


There is still a constant increase in comprehension up to deletion rate 10, but thereafter the pattern is less consistent. The easiest and intermediate texts show considerable differences between deletion rates 10, 12 and 14, but there is no agreement as to how they differ. Figure 4.3 compares the increase in restoration for the six

most difficult texts (grouped) with that of the six easiest texts (grouped). As before, there is a consistent gain in comprehension scores as the deletions increase in frequency, until deletion rate 10. Thereafter the curves go in opposite directions.

F I G U R E 4 . 3

Cloze scores: six most difficult texts vs. six easiest texts by deletion rate (Algeria)



If one draws the corresponding graphs for each text (Figure 4.4a and b), one produces a number of different patterns which have little in common apart from the fact that, with one exception (Text D1), all scores increase up to deletion rate 8. What happens is difficult to describe. It looks very much as if the less frequent deletion rates are giving unreliable results.

Previous studies have suggested that at least 30 deletions, and preferably 50, are required for a cloze test to give reliable results. In this study, deletion rates 12 and 14 have fewer than 30 deletions, and deletion rate 10 only occasionally achieves 30 deletions. To have achieved 50 deletions for rate 14 would have necessitated a text of at least 700 words, which would have meant that deletion rate 14 was, in effect, a different text from deletion rate 4 on the same passage (unless one deleted every fourth word for 700 words, which would have resulted in a somewhat tedious cloze test), since the content would be much greater. For this study it was decided to take texts of about 250-300 words in length, and keep the same length for each deletion rate, since it was felt important to have comparable texts. If two excerpts are made from one passage, one twice as long as the other, then two different texts have been produced. This assumes that the larger the content of one (i.e., the greater amount of reference to the same universe of discourse), the more different the two excerpts are, so that a point is reached when they cannot be said to be the same text. And if one were to have the same number of deletions per test, then the length of the text would have to vary, so that comparability of texts would, presumably, have been lost. It was felt worthwhile to run the risk of

having too few items for the less frequent deletion rates if one could gain the advantage of having the same amount of content for each deletion rate, since it seemed desirable to be able to claim that each test for any one text had the same overall reference.

With hindsight, it appears impossible to reconcile the desire to achieve a reliable result with the desire to have tests which are comparable in that they both refer to the same universe of discourse. Inevitably, a more frequent deletion system creates a different text from that created by a less frequent deletion system, even if the lengths are the same, simply because more words have been removed. And since, as a result of this pilot study, the deletion rates 10 - 14 now look like an interesting area for research it will be important to get reliable results, regardless of the length of the text.

Since no item analysis was carried out, normal internal consistency reliability estimates were not possible, so the only formula that might be useable here for the determination of reliability is the Kuder-Richardson 21. This formula assumes that all items are of equal difficulty. If they are not, then it will give a lower estimate of reliability than K-R 20, for example. However, it also assumes that all persons have attempted all items, which is clearly not the case here, due to the high number of items unanswered. It is, therefore, somewhat difficult to apply it to these data. Notwithstanding, calculations based on the K-R 21 formula, however inappropriate, were made, and they indicate that only deletion rates 4, 6 and 8, in general, achieve coefficients above .6, whereas 14, 12 and 10 are well below this. (See Table 4.8)

4) Finally, further tests (Table 4.9) were made for signifi-

cance of difference between means using the t-test for independent samples (since, if a subject had completed, say, Text A1, deletion rate 14, he obviously did not do Text A1, deletion rate 4). These results are consistent only in that deletion rate 4 is always significantly different from at least one other deletion rate (not necessarily, but usually, deletion rate 14). Apart from this there is a bewildering lack of agreement as to the effect that varying deletion rates has on "comprehension" or "restorability" of different texts. Thus for Texts F1 and E1, deletion rate 10 is significantly different from the rest, but the same is true for deletion rate 12 for Text F2. Worse, whereas on Text A2 virtually all deletion rates are significantly different from one another except for the relationship between deletion rates 8, 6 and 4, on Texts A1, C1 and D1 virtually all deletion rates give similar cloze scores.

Admittedly, in each case the number of subjects is low (between 14 and 26, average about 20), and one would expect occasionally to get significant results by chance somewhere among the 180 t-tests computed. Nevertheless, it looks as if the simple state of affairs indicated when one ignores all text differences is, in fact, not so simple when one looks at each text individually. In other words, from the data gained so far, one cannot make any generalisations about the effect of varying the rate of deletion from a text.

The only conclusion drawable at present is that the nature of the text used for the study - its style, readability, or some other variable - alters the effect of changing the deletion rate. The indications are that a person (non-native speaker) can expect to score

better on a cloze test made at deletion rate 6 than at deletion rate 4, and probably better on a test made at deletion rate 8 than at deletion rate 6. Where the levelling-off occurs, if at all, is uncertain, but probably around deletion rate 10. And, of course, whether this is true for a scoring method which accepts not only the exact word deleted, but also any contextually acceptable word, or even a word in the same form class as the deleted word, is not known.

4.7 Textual difficulty and cloze scores

Since cloze is expected to differentiate between texts of differing difficulty (hence its use as a measure of readability), it is not surprising that the cloze scores of the texts used in this study vary from text to text. Clearly cloze is a measure of some kind of text difficulty for non-native speakers of English. One problem is to see how this relates to other indices of text difficulty.

Readability formulae have been developed for use with native speakers of English, but little research has been carried out into their applicability to foreign learners. Thus, attempts at correlating text rankings based on cloze scores with rankings from a readability formula are, at best, questionable. However, for what it is worth, the Gunning Fog Index was applied to the texts used in the study, and the rankings compared with the cloze rankings, regardless of deletion rate (Table 4.10). The resulting Spearman correlation of .27 (Table 4.10b) was not significantly different from zero - i.e., no association was found between cloze rankings and a readability formula's rankings of the same texts. If one assumes that cloze, not Fog, is the better measure of readability for EFL students, then it appears that Fog is entirely

inappropriate as such a measure. This, moreover, is true regardless of the deletion rate used in the cloze test, since no deletion rate correlates significantly differently from zero with the Fog rankings.

Interestingly, the agreement among deletion rates as to the relative difficulty of these texts is low. Deletion rates 14 and 12 do not agree with any other deletion rate in the rankings, and deletion rate 10 only agrees with deletion rate 4. Closer agreement is achieved between deletion frequencies of every 8^{th}, 6^{th} and 4^{th} word, but even here it is far from perfect (the highest agreement reached is between deletion rates 8 and 6 at .89).

t-tests were calculated for differences between pairs of texts for those subjects who had taken one of each pair (Table 4.11a). Since one subject might have taken Texts A1 and C1 but not A1 and C2, the groups of subjects are different in each case. Since any subject who took, say, Text A1 did not take Text A2, t-tests for independent samples were also calculated to compare means for the two tests at any specific difficulty level (Table 4.11b). In this latter calculation, it was evident that those texts which were thought to be of approximately equal difficulty (at least as measured by Fog) were significantly different from each other, with only one exception (at level D). Thus the cloze test seems to be capable of distinguishing apparently similar texts (when differences in deletion rates are ignored). However, Table 4.11a shows that the cloze does not distinguish between texts which are apparently (according to Fog) very different in difficulty - for example, Texts A1 and F1, or A1 and D2.

Since, from Table 4.11a, it is clear that some texts were

not significantly different from others, it might be thought that a comparison of only those texts achieving significant differences from the ones above and below on the rank scale would yield a more meaningful relationship. This is because texts not significantly different from others could be ranked differently in relation to those texts by chance alone. However, when such texts (B2, C2, A1, F2, B1) were compared, no significant association was found between their cloze ranks and their Fog ranks.

Why should this be? Why should Fog (chosen mainly because of the simplicity of its application, but also because of its obvious similarity to other, perhaps better known, measures like Flesch that do not involve word frequency counts, which have less face validity for foreigners) be such a poor predictor of text difficulty for these students? Two reasons are suggested: one, that it is not linguistic enough - that is, that the length of sentence is not an adequate measure of syntactic complexity <u>for an EFL</u> student (however adequate it may be for native speakers); and two, that the word difficulty measure of words comprising three or more syllables is inappropriate in general for mature foreign learners of English who are, presumably, already fluent readers in their native language, and inappropriate in particular for speakers of French, for whom the ease caused by the familiarity of English polysyllabic words of Romance origin may well be greater than the difficulty occasioned by their polysyllabicity, and who may, in fact, find mono- and disyllabic words, which tend to be of Germanic origin, more difficult than tri- or polysyllabic words. Indeed, it seems doubtful that a syllable count, however valid for native

speakers, is an adequate measure of the lexical difficulty of a passage for EFL students.

Possibly, different measures of readability might be more appropriate to the estimation of difficulties for foreigners, although none are known to the author. Perhaps experienced teachers' ratings of the difficulty of these texts should be compared with obtained cloze scores to see if cloze is ranking these passages in an intuitively satisfying manner.

If one looks at the type of text found easiest and most difficult, it appears that the students found newspaper articles and prose fiction harder than other types of text, and non-academic essays easier than the rest. Yet, interestingly, these students were exposed far more frequently in class to newspaper articles and extracts from literary texts than essay-type texts. Presumably, degree of familiarity with a style does not help in the comprehension of any particular example of that style. The lack of familiarity with the content of the passage may have an effect, though it is not possible to substantiate this here, since no attempt was made to assess familiarity with content. Curiously, the text one might expect these students to be least familiar with - Orwell on mining - seems to have presented no outstanding difficulty. The more original use of words in prose fiction, combined with the presumed lower redundancy of literary text in general, may have been one of the causes of difficulty. Since such speculation cannot be confirmed, however, there seems little point in continuing along this avenue of enquiry, other than to point out that whereas two texts of Orwell, from the same set of essays, obtained widely different

cloze scores (C1, B2), two texts from the same short story by Rhys Davies were not significantly different (B1, D2, Table 4.11a), and that passages were no more closely related to other extracts from the same style than to passages from another style.

As has already been suggested, some texts are more alike than others, according to their cloze score, whatever Fog indices may indicate (Table 4.11). No one text is entirely different from the rest, although B2 (the easiest, according to cloze) is significantly different from ten of the other eleven texts. Similarly, no one text is more closely related to the rest than any other although both C1 and E1 are similar to six other texts (not the same texts in both cases). There is a tendency for the more difficult texts to be more similar than the intermediate and easy texts. (It is interesting to note here that if a text only achieves a cloze score of about 20%, it is less likely to be distinguishable from other, difficult, texts than if it had scored higher - in other words, there seems to be a base, or minimum, of about 20% comprehension/restoration for EFL students, below which texts do not go, and around which difficult texts tend to group.) This suggests that whilst it is important to consider text difficulty when constructing cloze tests, it is not the case that any two texts will give different cloze scores. However, since cloze does not correlate with Fog, it is useless to use the Fog Index as a predictor of the similarity of two texts. One cannot know whether two texts will give similar or different results until they have been tried out.

How does all this relate to the difference, or lack of difference, between deletion rates? Table 4.10b shows that differing deletion rates

do not always agree on the difficulty of texts. In particular, there is no agreement between deletion rates 14 and 12, and the cloze total ranking, or other deletion rates, and only slightly more agreement between deletion rate 10 and the rest. If this is not due to the unreliability of these particular deletion rates mentioned elsewhere, then it means that it is most important to take into account the deletion rates used for readability studies for foreign learners. That it is likely to be due to test unreliability is indicated by the results for deletion rates 8, 6 and 4 (all of which had an adequate number of items). These three deletion rates intercorrelate highly, and also agree with the ranking achieved by the cloze total. In other words, it does not matter whether one uses deletion rate 8, 6 or 4 as far as text ranking is concerned, but there is a possibility that a less frequent deletion rate than 10 changes the difference between texts. However, this may be due to the inherent unreliability caused by fewer items having been deleted.

As expected, cloze scores vary from text to text. Texts which seem to be more difficult to read get lower cloze scores than apparently easy texts. However, texts that appear to be easy for native speakers may be quite difficult for learners of English, and apparently more difficult texts may be relatively easier for them. The results of this study show that if the Fog Index is taken to be a useful measure of the difficulty of texts for native speakers, and the cloze scores to be a measure of difficulty for these foreign learners of English, then the nature of text difficulty is quite different for natives and foreigners.

One problem in this part of the study is that although an apparent range of text difficulty was achieved, as measured by a native-

speaker-intended formula, no one text obtained cloze scores, from these EFL students, above the "frustration" reading level identified by Bormuth and others for native speakers. However, even if this "frustration" level is applicable to non-natives, it is clear that texts can be differentiated at a low level of comprehension.

As far as this study is concerned, the main conclusion seems to be that texts are ranked in more or less the same order by frequent deletion rates, but that with less frequent deletion more research is needed to determine whether the increase in contextual constraint changes the relationship between texts. If the effect of text is the same regardless of deletion rate (i.e., if there is no interaction between deletion rate and text), then future studies into the nature of contextual constraint or the effect of changing the deletion systems need not expect widely different results from different texts.

If it is true that all deletion rates rank texts in the same order, then one can ignore text differences in deletion rate research. If, however, some deletion rates rank texts differently, then the type of text used, its difficulty level, and whether it is used with native or non-native speakers of English will probably all be relevant variables.

4.8 The relationship between cloze and a measure of English proficiency

Pearson Product Moment Correlations were computed to try to throw light on the following relationships:

a) Cloze total score & EPTB total & subtests (Table 4.12) (Table 4.15)

b) Intercorrelation of deletion rates (Table 4.13; Table 4.16)

c) Cloze deletion rates & EPTB total and subtests (Table 4.14; Table 4.17)

d) Each test (i.e., every text, every deletion rate) with EPTB Test 3

(Table 4.18)

e) Each test with EPTB Test 4 (Table 4.19) and total (Table 4.20)

f) Intercorrelation of texts (Table 4.21)

g) Cloze texts and EPTB total and subtests (Table 4.22)

4.8.1 Results

The results of the comparison between cloze total scores and EPTB are seen in Table 4.12.

TABLE 4.12

Relationship between total cloze score and EPTB (all students)

	EPTB Total	Test 1	Test 2	Test 3	Test 4
Cloze total score	+ .66	+ .34	+ .43	+ .55	+ .64
		$P < .01$			n = 139

A moderately high correlation (+0.66) is shown with the total measure of English proficiency, which is taken to mean that cloze is measuring a degree of language ability, but that either it measures an area of EFL proficiency not tapped by EPTB, or that cloze is not a language proficiency test by the standards of traditional tests (on the assumption that EPTB is a reliable and valid measure of that proficiency). The latter conclusion would appear to confirm Carroll's finding (Carroll et al, 1959) that cloze is not a measure of individual ability.

Comparison with the subtests reveals the fact that cloze is relatively unconnected with whatever is tested in the phoneme dis-

crimination test, Test 1, and the intonation test, Test 2 (.34 and .43 respectively). This is not surprising in view of Oller's claim that cloze is an integrative test, testing integrated skills, instead of discrete items or sub-skills. Since no dictation test was administered, it is impossible to confirm Oller's finding that cloze is most highly correlated with this "integrative" test. However, Test 2 is felt to test much more than just "intonation and stress", and would seem to approach a general listening test, in which case one might have expected a much closer relationship with cloze. Test 3 is itself a sort of cloze test, and so higher correlations were to be expected than were in fact achieved (.55). However, it is not a random deletion cloze, the deletions having been selected on syntactic criteria. Although there is clearly a reading component in this subtest, it is assumed by the author to be a test of grammatical relations. The cloze tests presumably tested more than the ability to predict grammatical relations, since the deletions included all possible form classes; yet the highest correlation between cloze and the subtests is found with Test 4 (.64), which explicitly tests grammar. One can only conclude that the cloze tests are testing to a considerable degree something called "grammar" which is more closely connected with the "grammar" tested by the discrete sentence items of Test 4 than that tested by the text of Test 3. This would appear to be an argument against Oller, since Test 3 is to be presumed more integrative than Test 4.

It was hypothesised that if cloze tested aspects of English language proficiency, a more frequent deletion from the text would relate more closely to EPTB than would a less frequent one. Also, the absence

of context (as achieved in deletion rate 4) would cause students to rely more on their grammatical abilities than their semantic abilities, whereas the presence of more context (as achieved by deleting only every 14th word) would enable students to use their semantic and discourse-rhetoric-hunting abilities. Therefore, deletion rate 14 would be much less closely related to Test 4 than would deletion rate 4, and the levels in between would show an increasing relationship as they neared deletion rate 4. It was further expected that the correlation between deletion rate and subtest would decrease as the subtest became less a test of integrative skills, and more a test of discrete skills. Table 4.14 shows that whilst the latter hypothesis was confirmed, since low or zero correlations were achieved with Test 1, the former hypothesis was not confirmed.

T A B L E 4 . 1 4

Relationship between cloze deletion rates and EPTB, total & sub-tests (all students)

Deletion rate	14	12	10	8	6	4
EPTB total	.43	.44	.50	.41	.43	.52
Test 4	.33	.50	.49	.45	.40	.48
Test 3	.31	.31	.34	.41	.39	.51
Test 2	.37	.25	.38	.20	.28	.29
Test 1	.24	.26	.29	.17++	.20++	.25

$p < .01$ Except ++ $n = 139$

++ = $p > .01 < .05$

There is no consistent change in the relationship between cloze and EPTB as the deletion rate changes although deletion rate 4 correlates markedly higher with EPTB Test 3 than do the other deletion rates, and deletion rate 14 correlates markedly less with EPTB Test 4 than do the other deletion rates. It does seem from this evidence, however, that whatever cloze measures, it is measured more or less equally by any deletion rate. It is noticeable that individual deletion rates correlate less with EPTB than does a total gained from a series of measures.

Just as the correlation of the total cloze score with EPTB increased from Test 1 to Test 4, so too the correlation of individual deletion rates and EPTB increases from Test 1 to Test 4. Usually the highest correlations (although still at a fairly low level - from .31 to .51) are achieved with Tests 3 and 4.

Further, it appears from Table 4.13 that the individual cloze tests are not measuring the same thing, since they have low intercorrelations, although the correlation with the total cloze score is approximately the same for each deletion rate (from .62 to .68). In view of the above, this is remarkable. Perhaps the answer is that the only thing each deletion rate has in common with the others is the measurement of EFL ability, and that this is only a small part of whatever these tests are measuring.

However, two factors may have invalidated the statistical analysis. Firstly, the assumption was made that the variation in textual difficulty was negligible, or at least could be ignored because a range of texts had been given, and that for any deletion rate, approximately

equal numbers of students had taken each text. It could be that the difficulty of text had not been cancelled out, and so further correlations were computed to investigate the effect of changes in textual difficulty on the relationship with EFL proficiency.

Secondly, blank tests were counted on the assumption that the student had found that particular deletion rate too difficult, and that zero was therefore a reflection of test difficulty. Some students left two or three tests blank, while scoring moderately or even very well on the remaining tests. This could have reduced the correlation. However, on correlating EPTB and cloze results only for those students who attempted every cloze test in their booklets, the deletion rate correlations remained about the same as for all students (Tables 4.15, 4.16 and 4.17). The differences were unpatterned, and there were as many cases of higher correlations as of lower correlations (5). For the correlation of deletion rate with EPTB, the coefficient was in general lower, with only three out of 36 substantially higher than data including blank tests, and with six correlations failing to achieve the 5% significance level. In all cases, however, the relationships remained the same - i.e., the correlation with deletion rate increases from Test 1 to the Total, the intercorrelations of deletion rates remained low, and the correlation with the cloze total remained moderately high at .61 to .74. Thus it seems reasonable to conclude that the effect of zero scores is about the same for each deletion rate, and that the difference between rates has been preserved. Further analyses will therefore ignore the effect of those students who only completed two or three of the cloze battery of six tests, whilst it must be conceded that the fact that this

happened is a weakness of the study.

In general, the number of subjects that had taken any one cloze test and EPTB was so low (around 10 per test) that few correlations of EPTB with individual cloze tests had a coefficient significantly different from zero (Tables 4.18, 4.19, 4.20). Fewer significant correlations were noted between cloze tests and Tests 1 and 2 of EPTB (11 and 14 out of 72) than between cloze and Tests 3, 4 and the Total (22, 23 and 27 respectively). This, of course, corresponds to what was discovered about the relationship between EPTB and the deletion rates in general.

Why should some tests show significant correlations, and others not? No one text achieves a consistently high correlation (Table 4.20).

T A B L E 4 . 2 0

Relationship between individual cloze scores and EPTB Total

Deletion rate/text	14	12	10	8	6	4
A1	NS	NS	+.61	NS	NS	NS
A2	NS	NS	+.80	NS	+.65	NS
B1	NS	+.64	NS	NS	+.86	+.76
B2	NS	NS	NS	+.74	NS	+.79
C1	NS	+.81	NS	NS	NS	NS
C2	+.63	NS	NS	NS	NS	+.70
D1	+.80	NS	NS	NS	NS	NS
D2	-.67	+.65	-.72	+.63	+.71	+.71
E1	NS	NS	+.61	+.73	+.69	NS
E2	NS	NS	+.89	NS	NS	+.67
F1	NS	NS	+.67	+.84	NS	+.53
F2	+.60	NS	NS	NS	+.78	NS

NS Not significantly different from zero

Although Text D2 correlates five out of six times significantly ($p < .05$) with the EPTB total score, one of these correlations, Text D2 at deletion rate 10, is high negative, whilst the rest are high positive. One can hardly be expected to place much trust in these figures, since it is intuitively unlikely that deletion rate 10 measures something entirely opposite from that measure by deletion rate 12 on the same text. It is presumably due to chance that those students taking Test D2 at deletion

rate 10 were poor in their English ability and good on cloze. Similarly, no one text achieves a consistent zero correlation with EPTB. Although Text B2 does not correlate with Test 3 of EPTB (Table 4.18), it does correlate significantly at deletion rates 12, 6 and 4 with Test 4 (Table 4.19), and at deletion rates 8 and 4 with the EPTB Total (Table 4.20). No one deletion rate has consistently high or low correlations with any part of EPTB.

In short, the results of the correlations are inconclusive in that they do not provide definitive proof of the thesis that deletion rates on some texts will have a greater relationship with English proficiency than deletion rates on other texts. If it is true that correlations between deletion rate and proficiency measures do not vary from text to text, it is perfectly valid to ignore textual differences for the purpose of examining the relationships between EFL proficiency and various cloze deletions.

The question that now arises is whether any text, because of its language, is a better predictor of EFL proficiency than any other. It might be expected that the comprehension of a difficult text (e.g., F1) involves greater EFL proficiency than the comprehension of an easy text. Since the relationship between deletion rates and EPTB is approximately the same for all deletion rates, it seems valid to ignore the deletion rate a student took on any given text, and so to regard, say, Text B2 at deletion rate 10 as equivalent to Text B2 at deletion rate 12 or deletion rate 4, for the purpose of this analysis.

In general, the intercorrelations between texts are fairly low (Table 4.21) with 25% not significantly different from zero. Text F2

seems to have the least relationship with the others (60% not significant) and Text F1 seems to have the most relationship. The most difficult texts tend to have a lower correlation with the other texts, and the easiest texts tend to have a higher correlation. The fact that the texts have low intercorrelations indicates that they have little in common and that therefore the type of text used in a cloze test has a great influence on the results obtained.

T A B L E 4 . 2 2

Relationship of text to EPTB, Total and subtests

EPTB/Test	Test 1	Test 2	Test 3	Test 4	Total
A1	NS	.45	.37	.37	.47
A2	NS	NS	.38	.49	.39
B1	.39	.45	NS	.37	.45
B2	NS	.27	.43	.66	.55
C1	.38	NS	.46	.46	.50
C2	.24	NS	.31	.34	.39
D1	.26	.29	.41	.50	.47
D2	NS	.25	.38	.36	.45
E1	.33	.47	.49	.59	.59
E2	NS	.24	.37	.33	.39
F1	.28	.41	.38	.51	.51
F2	NS	NS	.43	.44	.45

NS Not significantly different from zero

As expected, the correlation between text scores and EPTB scores (Table 4.22) increases as the subtest changes from aural to reading, and from involving discrete sub-skills to involving more general and global skills. However, again, the grammar test (Test 4) correlates more highly for virtually every text than does the modified cloze test (Test 3), and the correlation with the Total is more or less the same as the correlation with Test 4 for every text.

No one text has an obviously closer relationship with EFL proficiency than any other, although Texts B2, E1 and F1 do get somewhat higher correlations than the rest. Although there are no obvious differences between texts in their ability to predict scores on EPTB, the lower correlations tend to be with those texts that had a high cloze average score - i.e., the easier texts - and there is also a tendency for the more difficult texts to have a higher correlation with EPTB, especially with the total EPTB score. Thus, one might tentatively conclude that the ability to read apparently difficult text is more related to English proficiency than the ability to read an apparently easy text. It must, of course, be remembered that the texts provided here do not represent the extremes of ease and difficulty for texts, at least for foreign students, and it is conceivable that had much more difficult and much easier texts been used, clearer results might have been achieved.

That both the comprehension of difficult and easy texts, and successful completion of EPTB may (indeed, probably do) involve factors other than EFL proficiency, such as general IQ, verbal fluency, associational fluency, test-taking abilities, etc., is not denied. These factors simply cannot be isolated using the data presently available.

The object of this study has been twofold. Firstly, to see if any deletion rate or text or a combination of the two could be seen as the "best buy" in the prediction of EFL proficiency. The answer to this question has not been found. No deletion rate correlates consistently more highly with EPTB than any other, nor does the comprehension of any one text show a meaningfully closer relationship to EFL proficiency than any other. Only 9 out of 72 tests consistently showed significant correlations with EPTB, and of these nine no one is noticeably much better at prediction than any other.

The second aspect of the study, which, in a sense, is the first aspect from a different angle, was to see whether different texts or different deletion rates measured different aspects of EFL proficiency, or the same aspect in different amounts. No evidence was found for the notion that either texts or deletion rates measure different aspects of EFL proficiency as measured by EPTB, but some slight evidence was found to suggest that perhaps the more difficult texts are somewhat more closely associated with EFL proficiency than the easier texts. Very little evidence was found to indicate that more frequent deletion has a closer relationship with EFL proficiency than less frequent deletion. The evidence was by no means conclusive, and further studies would be necessary before even tentative conclusions could be made.

4.9 Summary and conclusions

The major results of this exploratory study into the effect of changing certain variables on cloze scores, with non-native speakers, are as follows:

1. There seem to be significant differences among cloze tests constructed by deleting words at different frequencies. However, these differences are not in any consistent direction.

2. Deleting every fourth word results in a cloze test which is always significantly harder than any less frequent deletion. In other words, to eliminate the effect of deletion frequency, there should always be at least five words of context either side of a blank.

3. A different analysis showed no significant differences between deletion rates 14, 12, 10 and 8. This suggests that there ought always to be at least seven words of context around each blank.

4. The graphs of cloze scores show a tendency for cloze scores to increase steadily from deletion rate 4 to deletion 10, after which the graph line levels out. This suggests that at least nine words of context are necessary either side of a blank before the effect of deletion frequency can be considered to have been eliminated.

5. There is an interaction between text difficulty and deletion frequency. When texts are not aggregated, there is no consistency whatsoever in the significant differences. On some texts all the deletion rates are the same, on others they are all different, on some texts only one deletion rate is different, whilst on others only one deletion rate is different from the rest although it is a different deletion rate from that on other texts.

6. There seems to be no relationship between cloze and Fog. Cloze discriminates among texts that Fog indicates are similar, and fails to discriminate among texts that Fog indicates are different. If Fog is a valid measure of readability for non-native speakers, then cloze is not.

If cloze is a valid measure of readability for non-natives, then Fog is not. However, it is possible that Fog is valid for native speakers, and cloze valid for non-native speakers, in which case one would conclude that the nature of readability for the two populations is different.

7. Cloze did not distinguish all the texts from all the others. It is difficult to say whether this is because cloze is insensitive to subtle differences, or whether the texts really were similar in difficulty.

8. There is little difference between deletion rates 8, 6 and 4 on the ranking of texts according to their difficulty. The less frequent deletion rates do not agree, however, and it is not clear whether this is because words were deleted infrequently or because the tests were less reliable.

9. Although there were variations in the correlation of different deletion rates with a measure of proficiency in English as a foreign language, the differences were small, and no consistent change in the coefficient with change in deletion frequency was observed. However, the intercorrelation of the different deletion rates was very low.

10. Text and deletion rate seem to interact on the correlation with proficiency but, again, no consistent pattern emerged. No "best buy" of combinations of text and deletion rate could be recommended.

11. Difficult texts tend to correlate higher with proficiency than do easy texts.

12. Cloze relates more to a test of grammar than to a listening test, or a phoneme discrimination test. This is contrary to previous research. The correlation with grammar was higher than that with a modified rational cloze procedure. No evidence was found that cloze is

more integrative than discrete point.

13. Even the overall correlation with proficiency was only moderate.

The variables studied do seem to have had an effect on cloze scores, but the picture is far from clear. To some extent, this is due to the design of the pilot study, where students took a large number of cloze tests in order for many different texts and deletion frequencies to be sampled. What is needed now is a closer study using more reliable cloze tests, where the subjects are able to complete their cloze tasks. This would involve giving them fewer cloze tests - probably only one - to complete, in order to ensure maximum performance on each test.

The extremes of deletion frequency have been shown to be, in the case of deletion rate 4, consistently different from and harder than less frequent deletion, and, in the case of deletion rate 14, consistently the same as more frequent deletion rates. Therefore, further study should concentrate on the intermediate rates: 6, 8, 10 and 12.

A reduced number of texts is also indicated, which should be different from one another according to a series of criteria, not only the Fog Index. Obvious differences in difficulty would enable clearer results to emerge, both for the measurement of proficiency and also for the interaction of difficulty and deletion rate.

A fuller study should also investigate the effect of the scoring procedure on cloze scores, and its interaction with the other variables.

Finally, the effect of these variables on the performance of non-native speakers needs to be compared with their effect on native

speakers, to see if there are qualitative differences which could throw light on the nature of the cloze task.

CHAPTER 5

The Design of the Main Study

Following on from the pilot study, it was decided to investigate three variables in the cloze procedure: text, deletion frequency and scoring procedure. It was hypothesised that these variables and their interactions would have an effect on cloze scores, both for native speakers of English, and for non-native speakers. It was hoped that by using the same tests and procedures with both populations, it would be possible to compare their performances. At the same time, it was hoped to be able to examine cloze as a measure of proficiency in English as a foreign language, and to explore the relationships between cloze, a traditional discrete-point test, and a more integrative test.

5.1 Hypotheses

The following are the null forms of the hypotheses the study was intended to test:

1a. There is no significant difference between cloze scores when deletion frequency changes.

1b. There is no significant interaction between deletion rate and text for easy, medium and difficult texts.

2. There is no difference in ranking of texts by different deletion rates.

   Subhypothesis: There is no difference in ranking of texts by different deletion rates when scored by different methods.

3a. There is no significant difference between exact and other scoring

methods.

3b. There is no significant difference between deletion rates when scored by exact word, and when scored by other methods.

4a. There is no difference between deletion rates as measures of proficiency in English as a foreign language (EFL).

4b. There is no difference between texts as measures of proficiency in EFL.

4c. There is no difference between scoring methods as measures of proficiency in EFL.

4d. There is no interaction between deletion rate, text and scoring method as predictors of proficiency in EFL.

Hypotheses 1, 2 and 3 would be tested with native and non-native speakers of English; Hypothesis 4 with non-native speakers only. The native speakers would be aged 14 to 15; the non-native speakers would be adult learners of English as a foreign language, and post-graduate foreign students of subjects other than English.

5.2 Design outline

The Algerian experiment had confirmed previous findings by other investigators that a deletion rate of less than five is always more difficult than deletion rates of five and above. The pilot study had also suggested that deletion rate 14 did not add any information not already provided by deletion rates 10 and 12. It was thus decided to abandon the fourth-word deletion rate, and the fourteenth, and to concentrate investigation on deletion rates 6, 8, 10 and 12, where results had indicated the asymptote of a negatively accelerated curve was to be found.

It was further proposed to limit the number of texts used, but to have a greater range of difficulty. The pilot study showed that, on the whole, the texts chosen, although similar in difficulty to texts used by the subjects in class, were probably too difficult, and in many cases the cloze was incapable of discriminating among them. It was decided to use three texts only - at the "easy", "medium" and "difficult" levels of readability. Since readability formulae, or at least the Fog Index, seemed incapable of predicting cloze difficulty for non-natives, it was decided to determine the difficulty levels of texts by a variety of techniques, including the judgments of experienced teachers of English as a foreign language (Section 5.3).

Five scoring methods were elaborated for use on the cloze tests. (Section 5.4)

There were thus three independent variables - deletion rate, text difficulty and scoring method; one dependent variable - the cloze score; and, for non-native speakers, one external criterion - a measure of proficiency in English as a foreign language. The design was a 4 x 3 x 5 factorial, and, since each subject would take only one cloze test, it was a straight full factorial with replication, orthogonality being assured by equal numbers of subjects in each cell. Random distribution of tests would ensure equivalence of cells, and, in the case of the non-native speakers, equivalent EFL proficiency, which would in any case be checked by analysis of variance.

5.3 Selection of texts

5.3.1 Texts

Eight texts were chosen to represent an apparent range of

difficulty for EFL students. The style common to all was "imaginative writing", since it was felt to be the only style one could assume all students, of whatever background, to be familiar with, and since expository material would seem to present problems of content and especially vocabulary difficulty. The texts were taken from readers and textbooks commonly used with foreign learners of English. The authors'/publishers' intentions as to recipients were noted, as was the approximate level of difficulty intended, as indicated by the blurb. In addition, one apparently very easy text was chosen to help the graders establish the low point of difficulty. The easy text was at the 500-word level (as defined by Oxford University Press and L.A. Hill). The other eight texts ranged from L.A. Hill's 750-word level, through stories simplified to the 2,000-word level of the General Service List, to a short story intended for native speakers and occasionally used by the author of this study with university level students of English as a Foreign Language (for details, see Table 5.1; for samples of the texts and the instructions given to raters, see Appendix B).

5.3.2 Measures

These texts were then assembled in a booklet in random order and given to 19 experienced teachers of EFL to rank in order of difficulty. These 19 raters were also asked to assess how difficult each text might be for a notional "upper intermediate" foreign student. The results are presented in Table 5.2. Raters were remarkably consistent in their judgments of the three easiest passages, and fairly consistent in their identification of the most difficult passages, but there was less agreement on the order of texts of medium difficulty.

Nevertheless, the coefficient of concordance (Kendall's W) was .88 ($p < .001$) which indicates high, although not perfect rater agreement, and a high reliability level for these ratings. It was thus considered appropriate to rank the texts in terms of the sum of the judges' rankings.

Several measures of text readability traditionally used on texts intended for native speakers of English were also applied to the passages to attempt to gain further information on the relative difficulties of the texts, despite the fact that the previous study had shown Fog to be a totally inadequate predictor of cloze difficulty. The results of the calculations, and the intercorrelations of all ten measures, are shown in Table 5.3.

T A B L E 5 . 3

Ten different methods of estimating text difficulty:

Ranks:

Criterion		1	2	3	4	5	6	7	8	9	10	Total	rank
Text	A	1	1	1	1	2	1	1	1.5	1	1	11.5	1
	B	8	7	7	6	7	7	6	6	7	8	69	7
	C	4.5	4	5	4	1	4	4	3	4	4	37.5	4
	D	2	2	3	2	6	5.5	2	1.5	5	2	31	3
	E	3	3	2	3	4	2	3	5	2	3	30	2
	F	9	9	8	8	9	9	9	9	8	9	87	9
	G	4.5	5	4	5	3	3	5	4	3	5	41.5	5
	H	7	8	9	9	8	8	8	7	9	7	80	8
	J	6	6	6	7	5	5.5	7	8	6	6	62.5	6

Kendall's Coefficient of Concordance $W = .88$ $X^2 = 70.4$ $p < .001$

1 = Publishers' intention

2 = Teachers' rankings

3 = FOG

4 = SMOG

5 = Coleman 1

6 = Coleman 2

7 = Dale-Chall

8 = Word frequency

9 = Flesch

10 = Teacher judgment of difficulty

The Fog, Smog, Flesch and Dale-Chall measures are well known. Fog, Smog and Flesch are essentially measures of word length (on the assumption that length is related to familiarity and difficulty of words) and sentence length (supposedly a measure of sentence complexity) whilst Dale-Chall, which includes a sentence length measure, measures presumed word difficulty by counting the number of infrequent words (frequency defined as appearance on the Dale-Chall word list of 3,000). The Coleman 1 and 2 measures (Numbers 5 and 6 in Table 5.3) comprise two formulae based, unusually, on cloze scores for the Miller-Coleman passages. Formula 1, basically a count of monosyllabic words, is said to predict 74% of the variance of the mean cloze score that would be gained by subjects similar to those used in the original investigation. Formula 2 is an extension of this, including a sentence length measure, and is said to predict 80.5% of the variance. The two formulae proved to be the worst predictors of teachers' ratings of difficulty of text for non-natives, but their validity as predictors of cloze scores of non-native speakers remains unknown.

Measure 8, word frequency, is simply the percentage of words not on the Dale-Chall list of 3,000 words. It is, as noted above, part of the Dale-Chall formula, but was used separately to see if the omission of "sentence length" would result in more or less valid prediction. The results seem to show that better prediction of teachers' gradings is gained when sentence length is also taken into account.

Measure 10 is distinct from teachers' rankings in that raters were asked to judge how difficult or easy a text was for upper intermediate students. The gradings were summed, and texts ranked accordingly.

In general, this measure correlates slightly less well than teachers' rankings with the other measures, except, interestingly, for the correlation with the publishers' intentions. In other words, the best way to find out how difficult publishers think texts are is to ask teachers how difficult they think the texts are.

The coefficient of concordance (Kendall's W) for all ten measures is .88 ($p < .001$), indicating that these measures have a great deal in common, and are ranking in substantially the same way, with reasonable reliability. The best measure appears to be Number 2 (teachers' rankings) which has the highest mean correlation (.91).

5.3.3 Decision

Text A had been introduced merely as a marker of the baseline of difficulty, but it was not intended to use it in the main study because it seemed to be idiosyncratic in style, with an abnormally high redundancy, obviously written for beginning learners of English. The next easiest text was Text D, and so this was chosen to be the "easy" text. The most difficult text was clearly Text F, and thus became the "difficult" text. The median text was somewhat more problematic, but since Text G, ranked 5, was most commonly considered of medium difficulty, it was selected as the "medium" text. Therefore, Texts D, F and G were used in the main study as the basis for "easy", "difficult" and "medium" cloze tests respectively.

5.4 Scoring Procedures

5.4.1 Review of procedures

As reported in Chapter 2, several investigators have found

that the any-acceptable scoring procedure is better for non-native speakers than the exact word method, although it appears to make no difference to the estimates of readability and the correlations with validating criteria with native speakers.

Other scoring procedures have been less frequently investigated, especially procedures giving credit for grammatical correspondence between the deletion and its restoration by subjects.

The original studies of cloze investigated only scoring procedures which gave credit to responses which were semantically related to the deletion, and did not consider procedures giving credit for grammatically related responses. In fact, despite the claim by Anderson (1972) that a "commonly investigated scoring procedure . . . is giving partial credit for responses of the same grammatical class as the deleted word", little evidence to substantiate this remark has been found. Marshall (1970), Moores (1967) and Odom, Blanton and Nunnally (1967) all used a form class score in their investigation of the language abilities of deaf children, but none of them investigate the procedure as such.

The earliest reference to a vaguely grammatical cloze score was found in Hafner (1964), who used a "GCIA" score, - i.e., allowing credit for responses which, although semantically incorrect, were grammatically correct. Unfortunately, no details of this procedure are given. He found that this GCIA score was a worse predictor of marks on a reading methods course (presumably, reading achievement) than the exact cloze procedure (correlations were .47 and .65 respectively), and that it was less closely related to measures of intelligence than the exact word score. More surprisingly, the GCIA score only correlated at .61 with the

exact cloze score.

Bormuth (1965b) classified his cloze responses as follows: EGC: exact word, grammatically correct; EGI: exact word, grammatically incorrect; SGC: synonym of deletion, grammatically correct; SGI: synonym, grammatically incorrect; UGC: unrelated semantically to deletion, grammatically correct; UGI: unrelated semantically, grammatically incorrect. His findings included the fact that scores based on grammatically correct responses correlated positively with his criterion of comprehension (reading achievement scores), but that scores based on grammatically incorrect responses correlated negatively or not at all with comprehension. He also found that the correlation with comprehension increased with an increase in the similarity of the meaning of responses to the deletion. (The correlations of comprehension with the scores were: EGC .82, SGC .64, UGC .55). Bormuth concluded that "a subject's comprehension of a passage is dependent upon both his ability to interpret sentence structure correctly, and to understand the content." He also suggests that the results indicate that the comprehension of a passage is incomplete when the cloze response is not the exact word. Further, the discrimination among passages was greatest with the EGC score - thus, the exact word method is best. Unfortunately, Bormuth did not sum the EGC, SGC and UGC scores and correlate the result with his criterion. It is conceivable that a grammatically correct score would correlate highly with comprehension. His results may not be applicable to non-native speakers - one does not expect native speakers to make many grammatically incorrect responses - but unfortunately he does not give details of his procedure, and so it is not clear what exactly

"grammatically correct" is. It could conceivably have a variety of meanings: being from the same form class or having the same grammatical function as the deletion, or having the same tense, number, etc. as the deleted word. It could even mean responses which do not violate grammatical constraints of the context, regardless of the grammar of the deleted item.

Oller (1972) investigated scoring methods with non-native speakers of English, but he did not use a pure grammatical score. Instead, he had several procedures which weighted differently acceptable responses, responses violating long-range constraints and responses violating short-term constraint. He discovered that a score based on any acceptable response (violating no contextual constraints), correlated best with his criterion (the UCLA ESLPE), better even than a score which had the following component weighting: 2 (exact words + acceptables) + long-range violations + short-range violations. In other words, no increase in validity was gained by allowing responses that violated constraints (although only partial allowance was made). However, what Oller does not highlight, but what emerges from the results, is that allowing even ungrammatical responses - "I goes" for "I go" - in the above scheme resulted in <u>higher</u> validity coefficients than the pure exact word score ($r = .82$ and $.75$ respectively). There is reason to assume, therefore, that unweighted form class scores - or other measures of sensitivity to grammar - may result in higher validity with non-native speakers than the exact-word-only score.

Anderson (1972) investigated four scoring procedures with non-native speakers of English (ESL) in Papua New Guinea, with partial

credits for certain types of response: 1) verbatim only; 2) synonym (verbatim, 1 point; a synonym, 1/2 point); 3) alternative response (exact, 1 point; each response that "made sense", was grammatically correct in terms of number agreement and fitted the syntactic pattern of the context, 1/2 point); 4) grammatical class score (exact, 1 point; "each response of the same grammatical class as the deleted word, regardless of number or tense", 1/2 point). He found that the four procedures were equally effective in discriminating between the three passages used, and all four procedures ranked subjects similarly. Because the reliabilities of most of the procedures were high, Anderson concluded that the exact word method is best (i.e., since all procedures seem to be measuring the same thing, the easiest method is to be preferred). Again, however, because each procedure was weighted in favour of the exact word method, the conclusions should be interpreted cautiously. What is needed is a comparison of simple unweighted scoring procedures, where any response is either correct or incorrect according to the criteria for that procedure.

5.4.2 Grammatical scoring procedures

It was decided to use the following three grammatical scoring procedures with both native and non-native speakers, in order to measure sensitivity to syntax.

1) Same form class: If a response is a member of the same form class as the deleted word, it is counted correct; if not, it is incorrect. Multi-word answers are incorrect; the semantic fit of the response is ignored.

Broadly speaking, traditional practice, Fries (1952) and Bormuth were

followed in determining form class membership. The class of function words was for this purpose subdivided into nine groups. The listing of classes used follows, together with the number or letter Fries (1952) attaches to each class:

Proper noun (1); noun (1); pronoun (1); verb (2); adjective (3); adverb (4); verb particle (4); co-ordinating conjunction (E); subordinating conjunction (J); "not" (C); "there" (H); preposition (F); determiner (A); auxiliary (G & B); intensifiers (D); question words (I).

However, inevitably, problems arise with this procedure, and arguments of the nature "Is this the same form class as that?" abound. For example, is "as" in the same form class as "like" and "than", or is it a preposition? Is "Standard 4" the same form class as "Standard Class"? Are "He was due to go" and "He was able to go" equivalent? Further problems were posed by sentences like "The shop window was broken", where "big" replaces "shop". Clearly the function of the two items is the same - modifier - and this function just happens to be realisable by different form classes. However, both seem equally predictable and valid grammatically thus the production of one rather than the other does not necessarily reveal differing degrees of sensitivity to syntax.

There is also the problem of degree of severity of error. For example, in the contexts 1) "He gazed dreamily at the Baptist preacher." and 2) "They fought in order to sit next to Monroe.", the replacement of "Baptist" by "yellow" (different form class) is of a different order from the replacement of "next" by "yellow" (again,

different form class). In fact, "yellow" in the first example, although from a different form class, is a more acceptable error than "beside" would be as a response to the second, although it is, presumably, from the same form class.

For these reasons, the following scoring procedure was also used:

2) Acceptable form class, same function: If a response was from a form class which was acceptable in the context of the item, it was scored as correct, provided that the response had the same grammatical function as the deleted item. Grammaticality of concord, number, tense, etc., was ignored. Answers of two or more words were incorrect.

This procedure ruled out replacing the noun in the first example below with a verb, as in the second:

1) He felt the cool night air on his back.

2) He felt the cool night blowing on his back.

since the function of air is head of the noun phrase acting as object of the verb felt whilst blowing is a predicate relating to night, and its use necessitates a reinterpretation of night as head noun, sentence object, rather than noun adjunct modifier of a nominal.

For this purpose, determiners and modifiers were regarded as performing separate functions. Thus, in the environment

He sold . . . . . horses.

"some" and "old" were not considered equivalent.

As a guide to this scoring procedure, the following examples are offered. In all four cases one alternative would not be regarded as being an acceptable replacement for the other:

very
The . . . . . old gentleman was walking down the street.
nice

and
The man . . . . . his dog can be seen in the distance.
with

when
He came . . . . . she went.
and

farmers
The . . . . . sold all their food.
they

Although some of the problems of Procedure 1 are solved by this procedure, some remain. One is the problem of the nature of errors alluded to earlier, which is only partly solved by Procedure 2, which would still allow "beside to" as an acceptable response for "next to" in the example above. Another is the ignoring of grammaticality. This procedure allowed as correct responses not only:

A birds was singing happily.,

but also

He had climbed out of bed, dressed, and running until he was tired.,

and

He had was accustomed.

This is justified by saying that the subject has correctly identified the need for a verb or a noun, etc., and has merely made a morphological error. However, this measure of the ability to identify the grammatical function of the deleted item, as determined by the context, needs to be supplemented by a third grammatical procedure.

3) Grammatically correct response: Any response which fits the syntax of the context is correct. It must agree in number, concord, etc., with

the environment and be from an acceptable form class, but it need not have the same function as the deletion, provided that the function it does have is appropriate grammatically.

This means that in certain circumstances co-ordinating conjunctions can be replaced by subordinating conjunctions, in other cases not, as in the examples below.

1) He had enough money . . . $\overset{\text{and}}{\underset{\text{so}}{\cdot\cdot}}$ . . . he bought a Rolls Royce.

2) John . . $\overset{\text{because}}{\underset{\text{and}}{\cdot\cdot}}$ . . . Mary went for a walk.

Semantic relatedness is, of course, ignored, as is semantic appropriacy. Multi-word answers are incorrect.

This procedure is presumed to be some measure of the subject's ability to respond grammatically, of his mastery of syntax, which is especially appropriate for non-native speakers. For native speakers, one would expect high or maximum scores with this procedure.

In addition to these grammatical scoring procedures the exact word method, and the "any acceptable word" method were also used:

5.4.3 Any acceptable word:

One of the main objections to the use of an "any acceptable" scoring method for cloze has been that it is particularly difficult to decide what is acceptable and what is not. Is it, for example, acceptable to replace "Mr Vaughan" (Text D) with "Mr Smith", although no "Mr Smith" has been mentioned before in the text, and will not be mentioned later? What one marker chooses to call acceptable may well be unacceptable to another marker, who may be using narrower criteria. It is possible that one judge considers stylistic infelicities to be

unacceptable, whilst another judge will find them acceptable.

A further problem of such a scoring procedure is whether any one judge is capable of being consistent - will he judge the same response from the same person in the same manner on two different occasions?

In order to investigate these two problems of marker inter-agreement and reliability, the following study was set up.

5.4.3.1 . The task

Ten native speakers of English, all experienced teachers of English as a Foreign Language, all studying for the M.Sc. or the Ph.D. in Applied Linguistics at Edinburgh University, were given a copy of a cloze test, uncompleted, and a list of responses to that cloze test made by the native speaker subjects of the main study. For each cloze test item, the marker was to judge whether the responses provided were acceptable or unacceptable. No further guidance as to the nature of acceptability was given. An unacceptable response on the list was to be underlined, the acceptable responses to be left untouched.

The resulting list of acceptable and unacceptable responses for each judge was used as the basis of a scoring key for a computer scoring procedure. Computer scoring ensured 100% objectivity of scoring. The scores produced - hereinafter referred to as the judged scores - were taken to be the judges' criterion scores.

After at least one month had elapsed, the same judges were given a set of 30 cloze tests booklets, completed by the native speaker subjects of the main study, and were asked to mark them for acceptable responses. If acceptable, a response scored 1; if unacceptable, a response scored 0. All the investigator required from the judges was

the total score of acceptable responses for each individual.

It was then possible to compare marked scores for different markers, and judged scores for different judges, thus gaining measures of inter-marker and inter-judge agreement. It was also possible to compare the marked score of each marker with his judged score to produce a measure of marker reliability.

5.4.3.2 The text

The text used for this study was the medium difficulty text used in the main study. Preliminary investigations involved giving all three (difficult, medium and easy) texts to all judges, with the respective lists of responses (provided by the native speaker subjects) and asking for judgments of acceptability as outlined above. The different texts were then compared for amount of marker agreement, and it was discovered that both the easy and the difficult texts had a high proportion of responses which were judged as either entirely unacceptable or entirely acceptable by the judges, whereas the medium text had the highest number of indeterminately acceptable responses, in that fewer judges agreed on the acceptability of responses. So as not to bias the results by choosing a text on the acceptability of whose responses most of the judges agreed, the medium text was chosen for the investigation.

5.4.3.3 Follow-ups

A further study resulting from this investigation was carried out. This study involved seven of the previous native speaker judge/markers, and seven non-native speakers of English also as judges. All of the latter were students of Edinburgh University studying for the M.Sc. in Applied Linguistics. The countries of origin of these non-

native speakers were: India, Pakistan, Hong Kong, Malaysia, Brazil, Italy and Nigeria.

Using the same cloze test, a set of cloze test booklets, this time completed by the non-native speakers of the main study, was given to the judges to mark for acceptability.

It was possible to investigate the amount of marker agreement on non-native speaker responses, and to compare the agreement of native and non-native speaker markers. It has been suggested in the literature that whilst native speakers may well agree on acceptable responses, non-native speakers will not agree on the acceptability of responses, neither among themselves nor with native speakers.

Overall, including the follow-up study, the following measures are possible:

1. Native speaker judges/markers
   a) reliability of mark-remark (judged score - marked score inter-correlations) for each judge/marker, based on native speaker responses
   b) agreement, by judge, with all other judges of native speaker responses (judged score intercorrelations)
   c) agreement, by marker, with all other markers, of native speaker responses and non-native speaker responses (marked score inter-correlations)
   d) overall agreement on judged scores (Kendall's W), native speaker responses
   e) overall agreement on marked scores (Kendall's W), native and non-native speaker responses

2. Non-native speaker markers

   a) agreement, by marker, with all other non-native markers, of non-native speaker responses (marked score intercorrelations)

   b) overall agreement on non-native speaker responses (Kendall's W)

3. Comparison of native - non-native speaker markers

   a) native vs. non-native overall agreement on non-native responses

   b) native - non-native speaker marker agreement on non-native responses

5.4.3.4 Results and discussion

For convenience, the tables of results are labelled and numbered in exactly the same way as the above listing of possible measures, preceded by the numbers 5.4.

An examination of these tables gives rise to the following comments and conclusions.

TABLE 5.4.1a

Native speakers: Reliability of mark-remark (correlation of judged score and marked score, for native speaker responses)

	Pearson Product Moment
Judge 01	.93
Judge 02	.95
Judge 03	.96
Judge 04	.97
Judge 05	.95
Judge 06	.96
Judge 07	.96
Judge 08	.97
Judge 09	.98
Judge 10	.97

n = 30

TABLE 5.4.1b

Native speakers: Agreement, by judge, with all other judges, of native speaker responses (judged score intercorrelations)

Pearson Product Moment

Judge	02	03	04	05	06	07	08	09	10
01	.98	.94	.97	.95	.96	.96	.96	.97	.97
02		.95	.97	.96	.97	.97	.98	.96	.97
03			.93	.96	.94	.95	.98	.96	.93
04				.96	.98	.97	.97	.97	.99
05					.97	.97	.97	.97	.96
06						.98	.97	.98	.98
07							.97	.98	.98
08								.97	.97
09									.97

n = 30

1) It is quite clear that native speaker judges are capable of judging and marking cloze tests (filled in by native speakers of English) for acceptable responses, and this they do with a high degree of reliability (Table 5.4.1a). No reliability coefficient was lower than .93. Not only are the subjects given more or less the same score the second time round, but they are also given almost exactly the same rank relative to the other subjects. It is clear that this is true for all the native speaker judges, not just some of them.

2) Native speakers agree with each other to a remarkable degree as to the acceptability of responses. Despite the fact that the medium text was chosen since it showed the greatest amount of marker disagreement on the acceptability of individual responses, the actual scores produced by the acceptable procedure by different markers intercorrelate highly (from .92 to .98) (Kendall's W = .91). In other words, what disagreement there is is trivial, and confined to odd (i.e., deviant) and infrequent answers. What disagreement there is therefore barely affects the overall score for individuals.

3) The somewhat higher intercorrelations for judged scores for native speakers (from .93 to .99) could be due to two things: a) the fact that the judged scores were computer marked, and are thus 100% objective, or b) the lower objectivity of the marked scores (i.e., markers gave credit for one response for one subject, but not for another subject). Nevertheless, the difference between the marked score intercorrelations and the judged score intercorrelations seems trivial, and hardly invalidates the general conclusion that native speakers can reliably score cloze tests by the "any acceptable

word" procedure, and that they agree very closely with each other on the scores they give to native speaker subjects.

4.

T A B L E 5 . 4 .1e

Overall agreement, native speakers, marked scores

1) marking native speakers

Kendall's W = .91

average Spearman rho = .90

2) marking non-native speakers

Kendall's W = .95

average Spearman rho = .94

There is essentially no difference in the amount of agreement as to the acceptability of non-native speaker responses compared with the agreement on native speaker responses (Table 5.4.1e). A native speaker marker is just as capable of marking non-native speaker responses as he is of marking native speaker responses.

5. TABLE 5.4.2a

Agreement, by non-native speaker/marker, with all other non-native markers of non-native speaker responses.

Pearson Marker	31	32	33	34	35	36	37
31		.95	.89	.94	.94	.80	.94
32			.88	.97	.96	.86	.94
33				.89	.91	.84	.92
34					.96	.88	.92
35					.89	.94	.94
36							.87

n = 30

TABLE 5.4.2b

Overall non-native speaker agreement on non-native speaker responses

Kendall's W: .93

Average Spearman rho = .92

Non-native speaker markers have a high amount of agreement amongst themselves as to the acceptability of cloze responses. Intercorrelations range from .80 to .97. (Table 5.4.2)

6. TABLE 5.4.3b

Agreement of non-native speaker marker scores with native speaker marker scores, on non-native speaker subjects' responses.

Pearson Product Moment

Native speaker/ non-native speaker		11	12	13	14	15	16	17
	31	.93	.93	.86	.90	.94	.93	.96
	32	.94	.94	.90	.89	.94	.94	.95
	33	.94	.92	.84	.95	.91	.93	.93
	34	.95	.94	.94	.90	.95	.95	.96
	35	.96	.96	.91	.92	.96	.96	.96
	36	.90	.91	.85	.89	.91	.92	.89
	37	.95	.96	.88	.96	.98	.96	.96

n = 30

Non-native speakers agree in all essentials with native speakers as to the acceptability of responses. Or rather, if there is any disagreement, it has little or no effect - the subjects' scores remain virtually the same.

The general conclusion, then, must be that contrary to common opinion and supposition, a high degree of agreement can be gained by native and non-native speakers of English as to the acceptability of cloze responses provided by both native speakers and non-native speaker subjects. The "any acceptable" cloze scoring procedure is thus perfectly feasible in non-English speaking countries as well as in English-speaking countries.

Moreover, because of the high agreement amongst markers, it is unnecessary to have a battery of judges deciding on the acceptability of responses. One judge should be adequate, since it has been shown that all judges are capable of scoring for acceptability reliably and in a manner which agrees with what other judges would have done.

In view of the large number of cloze tests used in the main study, and the high number of different responses provided by the subjects, it would be highly impractical to engage a panel of judges to decide on acceptable responses. This short study showed that this is not only impractical but unnecessary. The judgment of acceptable responses, in preparation of a computer scoring key - which ensures 100% internal reliability - was therefore undertaken solely by the author. As a final check on the validity of this procedure, however, when the scoring of Test M12 had been completed for non-native speaker subjects, the results were compared with the scores produced by the 14 native and non-native speaker judges, and the correlations are tabulated in Table 5.5.

TABLE 5.5

Correlation of author's scoring (SEMAC = Semantically Acceptable Score) with non-native and native speaker markers, individually, by subgroups, and overall.

a) Pearson Product Moment

1) with native speakers

	11	12	13	14	15	16	17	Combined
SEMAC	.98	.97	.94	.95	.98	.98	.98	.99

2) with non-native speakers

	31	32	33	34	35	36	37	Combined
SEMAC	.93	.96	.94	.96	.97	.89	.97	.98

3) with overall score, based on 14 markers

r = .99

The intercorrelations were uniformly high (from .89 to .99), both with individual markers and with the combined mark from 14 markers, so it can be safely concluded that this author's judgment of acceptability of responses corresponds very closely, for all practical purposes, to the judgments of other educated native (and non-native) speakers of English. The "semantically acceptable" procedure used in the main study can thus be considered a valid procedure.

5.4.4 Summary

In addition to the exact word scoring procedure, four other scoring procedures were investigated: 1) any semantically acceptable word, 2) any grammatically acceptable restoration, 3) any restoration from an acceptable form class with the same grammatical function as the deletion, and 4) any restoration from the same form class as the deletion.

5.5 Administration procedure

Having selected the three cloze texts of differing difficulty, it was then necessary to submit them to four deletion procedures.

The first two or three sentences were left intact as a lead-in, and then, counting from the n^{th} word (n = 29 for the easy (E) text, 31 for the medium (M) text and 32 for the difficult (D) text), every

6^{th}, 8^{th}, 10^{th} or 12^{th} word was deleted, leaving a total of fifty deletions and giving a total of four tests per text and twelve tests in all. This figure of 50 is held by Bormuth and others to provide enough items for a reliable result to be achieved, and it is also thought to provide an adequate sample of the text for readability estimates. When fifty words had been deleted from the passage, the test was terminated at the end of the sentence then in progress. Each word was replaced by a fifteen-space-long line, preceded by the number of that item in the text. Thus the first word deleted was replaced by 1_____, the second by 2_____, and so on. This format was chosen because it was felt to be less disturbing of the reading process for the subjects to fill in deletions in the passage, rather than on a separate sheet of paper, or even in the margin of the same sheet, however convenient these might be for data processing. Furthermore, subjects would be able to revise the text and their restitutions as a whole, seeing the restorations in context.

Each of the resulting twelve tests was prefaced with a one-page instruction sheet containing the standard Bormuth instructions (Bormuth, 1964b), which were slightly altered to allow for the fact that both parts of hyphenated words, whether free or bound forms, were deleted, and which included a short four-item example. (Samples of the instruction sheet and the 12 tests are found in Appendix D.)

The tests were arranged in sets of twelve, in sequence, beginning with the D text at deletion rate 6, followed by the M text at the same deletion rate, the E text, and so on through the deletion rates. This was done to prevent cheating from one's neighbour, so the booklets

were distributed in this fixed order. It was assumed that subjects had distributed themselves randomly in class, there being no reason to assume that subjects of particular linguistic abilities would always position themselves in regular patterns, thereby countering the randomness of the booklet distribution. Each subject took only one of the possible 12 tests, in accordance with the experimental design; viz., a full factorial, with replication, subjects being independent.

Subjects were given as much time as required to complete the tests; nevertheless, some subjects either did not complete their test or restored less than half the items. These tests were removed from the subsequent analysis, the assumption being that their poor performance was due to poor reading abilities.

Obviously the words deleted by the above deletion system were different for each deletion rate, with maximum overlap between deletion rates 12 and 6, and some only between all deletion rates. It could therefore be argued that any difference that shows up between deletion rates is not due to the rate of deletion - i.e., to the number of words between deletions - but rather to the nature of the deletions. Conceivably one deletion rate could remove a greater percentage of content words than another, and thus gain a lower cloze score.

Bormuth (1964c) reports on an investigation into whether any one deletion form was equivalent to any other possible form at the same deletion rate - i.e., whether deletion rate 5 deleting the 5^{th}, 10^{th}, 15^{th} . . . words gave the same cloze score as the same rate starting at the 6^{th} word, or similarly for forms starting at the 7^{th}, 8^{th} and 9^{th} words. He discovered that even when there were 50 items in the test,

significant differences still existed between the forms for eight out of the twenty passages used ($p < .05$), and concluded that all possible patterns should be used when investigating any particular text by means of the cloze.

In the present case, this would mean six different forms at the 6^{th} deletion rate, eight at the 8^{th} rate, and so on, giving 36 tests for four deletion rates on one text alone. This would presumably solve the problem if the results for each form were averaged to give a score for each deletion rate. Whilst adding the scores for matched subjects from different forms may be valid for readability studies, it seems less valid for studies of comprehension and language ability, since, in effect, subjects on different forms would have been exposed to different texts.

In any case, 36 tests per text would be well beyond the scope of this study, whose purpose was to investigate the claim that varying the deletion rate has no great effect on the cloze score. This claim itself ignores the fact that different words are deleted by different rates. However, some attention could be given to the problem by looking at those items where overlap occurs - e.g., the 48^{th} word of text is Item 4 at deletion rate 12, Item 8 at deletion rate 6 and Item 6 at deletion rate 8. One would look to see if the scores for different amounts of context on either side of the deletion are in fact the same (as MacGinitie claimed) or different.

Details of the administration of the proficiency measures to the non-native speaker subjects are given in Chapter 7. Similarly, details of the subjects used in each part of the study precede the results of the relevant part - i.e., Chapter 6 for native speakers, and Chapter 7

for non-native speakers. For ease of comparison of native and non-native speaker subjects, Chapter 7 duplicates Chapter 6 but with different subjects. Those aspects of the non-native speaker study peculiar to itself are therefore dealt with separately in Chapter 8.

5.6 Measures of proficiency in English as a foreign language

With non-native speakers, it was decided to use two measures of EFL proficiency: a standardised test of traditional design, the English Language Battery, and two "integrative" dictation tests. The English Language Battery is described in Section 5.6.1, whilst Section 5.6.2 presents, at some length, a review of the literature on dictation, especially if its relationship with the cloze procedure, and a description of the tests used in this study.

5.6.1 The English Language Battery

The following description of the English Language Battery (ELBA) is taken from the test manual:

> "The English Language Battery (ELBA) is a proficiency test English as a foreign or second language. Its primary purpose is to distinguish students who have sufficient command of English to pursue their studies in institutions of higher education where the language of instruction is English, from those who will experience varying but serious language difficulties. In difficulty level it is suitable for good intermediate and advanced learners."

The test has two parts, Listening and Reading, of multiple-choice format. Part One (Listening) has four tests: Sound Recognition, Intonation, Stress, and Listening Comprehension, of the following nature:

1. Sound Recognition (100 items). One English word is read out on tape. This has to be matched to one of three English words printed on the answer sheet.

   e.g., /bit/ is heard on tape, and
   on the answer sheet is seen

   bit bet bat
   ( ) ( ) ( )

2. Intonation (10 items). A short sentence is printed on the answer sheet, together with three alternative interpretations of the speaker's intention. The sentence is read out on the tape with a distinctive intonation pattern.

   e.g., He speaks fluent Hindi. This probably
   a ( ) is a request for information.
   b ( ) expresses great surprise.
   c ( ) is a straightforward statement.

3. Stress (10 items). A short sentence is printed spaced into syllables. The same sentence is read on tape with the tonic located on a particular syllable. The student has to indicate the syllable on which he thinks the sentence stress falls.

   e.g., There's a let - ter for you.
   ( ) ( ) ( ) ( ) ( ) ( )

4. Listening Comprehension (30 items). The items consist of very short exchanges between two people. The first remark is recorded on tape, and is not written down. What is written down on the answer sheet are various alternative continuations or answers to the spoken remark, and the students have to choose the one that is most likely.

e.g., "Will you come with us on a picnic tomorrow?" (on tape)

a ( ) I'd like to if the weather is decent.

b ( ) When do the summer holidays start? (Printed on answer sheet)

c ( ) I'm sorry, I was too busy.

d ( ) I haven't seen you for ages.

Part Two (Reading) has three tests: Grammar, Vocabulary, and Reading Comprehension, of the following nature:

5. Grammar (50 items). Most of the items concern the choice of the most appropriate grammatical alternative in a given context.

e.g., ... of rain in Scotland.

a ( ) It's a lot

b ( ) This lot

c ( ) There's a lot

d ( ) It's lots

6. Vocabulary (50 items). No context is used for single words, and only a minimal amount for compound lexical items.

e.g., entertaining

a ( ) amusing

b ( ) laughing

c ( ) meddling

d ( ) generous

7. Reading Comprehension (20 items). Four short passages for comprehension which are supposed to represent the sort of material students might be expected to be able to process are followed by a series of multiple-choice questions, aimed largely at tapping

inferential skills.

Reliability

The Kuder-Richardson (21) reliability estimates reported in the manual are .97 for the total test, .93 for Part One, and .96 for Part Two.

Validity

The internal validity coefficients are reported as follows for a group of 320 non-native first-year students from Scottish and English universities and colleges of education.

	1	2	3	4	5	6	7	8	9	10
1 Sound										
2 Intonation	.56									
3 Stress	.58	.49								
4 List. Comp.	.72	.68	.50							
5 Part I	.93	.73	.69	.89						
6 Grammar	.68	.68	.40	.77	.77					
7 Vocabulary	.63	.52	.26	.69	.68	.78				
8 Read. Comp.	.59	.58	.46	.69	.70	.67	.62			
9 Part II	.71	.63	.39	.79	.79	.93	.93	.80		
10 ELBA Total	.85	.71	.56	.89	.94	.90	.87	.79	.96	

The external validity correlations were calculated for two criteria:

1) end of term examination in English language, 3 months after testing
2) teacher ratings, made at the same time as the testing

The coefficients for 1) are reported as being around .80, whilst those for 2) range from .61 to .91, depending on the subgroup

being used. Only one correlation is reported with the other major test of English as a Foreign Language used in the United Kingdom - the English Proficiency Test Battery (EPTB, or Davies Test), for which the coefficient, based on 59 subjects, was .68.

It would appear from these figures that the ELBA is a reliable and valid measure of proficiency in English as a foreign or second language, and is thus suitable for use in this study.

5.6.2 Dictation tests

In view of certain experimental results obtained by Oller in his investigations into the cloze procedure and dictation (see Chapter 2 for a brief introduction), it was decided to investigate the dictation further.

5.6.2.1 Review of the literature

What follows is a comprehensive review of attitudes toward dictation and experimental results gained in its use. This is intended as an introduction to the problems presented by dictation, and to the actual design of this study.

The following quotation from Lado is typical of the testing experts' attitude to dictation for the last fifteen to twenty years:

> "On critical inspection it appears to measure very little of language. Since the order of words is given by the examiner as he reads the material, it does not test word order. Since the words are given by the examiner, it does not test vocabulary. It hardly tests aural perception of the examiner's pronunciation, because the words can in many cases be identified by context if the student does not hear the sounds correctly . . . . Spelling and a few

matters of inflection and punctuation can be tested through dictation, but the complicated apparatus of dictation is not required to test these matters." (Lado, 1961)

David Harris said, "As a testing device . . . dictation must be regarded as generally both uneconomical and imprecise." (Harris, 1969)

Wilga Rivers, whilst pointing out that dictation could be a useful technique "for verifying whether students have learned to make certain discriminations among sounds", and that, at an advanced level, it can be used as a "test of the student's knowledge of combinations of letters which traditionally represent specific sounds, and as a test of his knowledge of structural elements, particularly those of a morphological nature", concludes that it presents many problems for the tester. Typically, students do not pay attention to the meaning of what they are writing, nor to the way the segments being dictated fit into the whole passage. It also tests auditory memory, and probably even temperament, in which case "it cannot be considered a valid test of listening comprehension alone", and is probably best used only as a teaching exercise (Rivers, 1968).

The general objection on the part of the experts to dictation seems to have been that whatever areas dictation does test are better (i.e., more reliably and validly) tested in other manners, and that there is a great deal of wasted data collected by a dictation.

Although dictation has continued to be used as a teaching device, there seems to have been a general agreement that it was an unrespectable testing device, so that even those who went into print to

advocate its use in class never went so far as to suggest that it might be used in a test. About as far as people dared go was to suggest, as Rivers did, that

> "it can be used as part of a group of tests aimed at determining all-round skill in handling the language, but it is doubtful whether it reveals anything which has not already been identified by other tests in a well-designed battery - in which case its contribution to the result may be considered largely redundant."

It was only rarely that dictation was defended in front of the experts. H.A. Cartledge (1968) defends dictation for its face validity and relevance (stenographers need dictation), and maintains that it gives practice in oral comprehension, since by transferring from spoken to written language we prove we have understood exactly what has been said (a claim that Lado et al. would disagree with). He also claims that dictation involves more than spelling abilities, that it "obliges students to contextualize and discriminate"; it is "an excellent way of assessing a student's grasp of current speech". Unfortunately this fulsome praise of dictation's virtues is based on virtually no evidence other than the anecdotal:

> "If I write from dictation 'I <u>read</u> the letter and underlined the important parts of it with a <u>red</u> biro', I use my knowledge of English to distinguish between the two uses of / red /. If I am unaware of the need for different spellings, I may not do this correctly, but if I am aware of it, the context of the words obliges me to use the appropriate spelling in each case."

This is, of course, true, but it is also true that one can be "unaware

of the need for different spellings", and still understand spoken English. Whatever it is, dictation is clearly not only a test of listening comprehension, which is why it lurked in disfavour for many years. It used to be axiomatic in language testing that each test item should be unambiguously testing one skill or subskill.

Now that integrative and global tests have become respectable, a reconsideration of the nature and usefulness of dictation has become inevitable, and its cause has been taken up in recent years by Oller and others. Oller points out that he was not the first to see the virtues of dictation. Most of the advocates of dictation he mentions, however, were chary of suggesting that it be used as a language test. Sawyer and Silver (1065) and Finocchiaro (1958) emphasise the usefulness of dictation as a classroom technique only. Rebecca Valette (1964) investigated the use of the "dictée" in French lessons, and discovered that it was a good predictor of overall language proficiency. She also discovered that practice in dictation, which improved performance on the dictation test, also reduced the ability of dictation to predict language proficiency. Practice appeared to result in "greater awareness of the written language" as measured by her writing test, but it did not produce better scores on the grammar subtest than lack of practice, and it resulted in significantly lower scores on the oral comprehension test. She concluded that "proficiency in dictation does not . . . imply proficiency in other aspects of French language learning." In her handbook, Modern Language Testing (1967), she mentions the disagreement amongst specialists as to dictation's usefulness as a language test, and concedes that the "art of taking dictation is a specialized skill". She

avoids controversy by saying that "whatever dictation lacks theoretically, dictation scores in practice correlate very well with overall achievement." She neither advocates nor condemns its use, she simply points out that it is used and that further research is necessary.

This is different from the support Oller gives the use of dictation. His goal is unashamedly no longer discrete items testing discrete skills, but a global test testing overall competence and proficiency. In Oller (1971) figures are produced which, it is claimed, show dictation to be the best single measure of the totality of English language skills being tested on the UCLA ESLPE Form 1. In this study dictation correlated highly (.86) with the ESLPE Total score. However, there are several points to be made about his data, which he does not mention. 1) All the subsections of the test correlate highly with the Total, only Phonology achieving a coefficient of less than .77. 2) Composition, not dictation, has the highest correlation with the Total. 3) The section weightings are not given. 4) Dictation correlates most highly with Composition and Vocabulary, one of which is presumably integrative, and the other discrete; and least with Phonology (.57), which is surprising, since both it and dictation deal with listening. It is not enough to claim, as Oller does, that dictation is integrative, and Phonology discrete, because of the correlation with Vocabulary mentioned above. 5) No details of how the dictation was administered and scored are given. It is impossible to say that spelling is not the major element in what is tested by dictation if one does not provide evidence that spelling was ignored in the scoring system. Nevertheless, Oller claims that previous authorities have been refuted

by this evidence, that dictation correlates more highly with every other part of the test than does any other part, and thus that dictation is the "best single measure of the totality of English language skills being tested".

Some of the above criticisms are answered in Oller and Streiff (1975), where because of similar criticism from Rand (1972), he discovered errors in his figures and reworked them. More importantly, he recognised that the differential weightings of the part tests would affect the correlations. (For the record, these were Vocabulary, 20; Grammar, 25; Composition, 25; Phonology, 15; Dictation, 15.) The reworked correlations give increased correlations between dictation and the other parts. With the Total it goes up from .86 to .94, with Vocabulary from .67 to .72, with Grammar from .64 to .65 and with Composition from .69 to .72. Only with Phonology is there no change. Moreover, the point made above about the correlation of the Total with Composition is no longer valid. Dictation now has clearly the highest correlation (.94 versus .85). Also, the correlations of dictation with the other sections are higher (.85) than any other section with the remaining sections. It looks very much as if Oller's thesis is borne out by the evidence. However, still unexplained are the following points: 1) Why does dictation correlate just as highly with Vocabulary (discrete) as with Composition (integrative), and least with Phonology? 2) Although Oller gives details of the administrative and scoring procedures, it is clear from the latter that spelling errors are considered equivalent to phonological, grammatical and lexical errors. From a maximum possible of 15 points for dictation, one quarter-point is deducted (and only one

quarter-point per word, regardless of how many errors were present in that word) for clear errors in spelling ("shagrin" for "chagrin"), phonology ("long hair" for "lawn care"), grammar ("it became" for "it becomes") or choice of wording ("humanity" for "mankind"). Since no details are given of the relative frequency of spelling, phonological or other errors, one still does not know whether the scores were largely made up of spelling errors, in which case the test was a spelling test. 3) There is no explanation for the fact that Vocabulary, Grammar and Composition all had approximately equal correlations with the remaining part-tests (albeit higher than Phonology and lower than dictation). Yet this is surely worthy of at least a comment, since Vocabulary and Grammar are to be presumed discrete tests, and Composition an integrative test.

Of the studies investigating the relationship between the cloze procedure and other measures of EFL proficiency, Darnell's study (1968) was the first to report on the relationship between cloze and dictation. With a sample of 40 college students, he found that the highest correlation for dictation was with the Gates Reading Survey (.78), and that the next highest was with the clozentropy test (.63). Lower correlations of .59 and .48 were obtained with the Lado Oral Comprehension Test and an oral interview respectively. Unfortunately no details were given of the dictation other than that it was marked on a 100-point scale, marks being deducted for mistakes in grammar, punctuation and spelling. Darnell himself makes no comment on the nature of the dictation or its correlations with the other tests, other than to say that the correlation with clozentropy justifies further investigation of clozentropy. It is, however, interesting that dictation seems to be more closely related to reading

tests than to tests of oral ability, regardless of the "integrative" or "discrete" nature of either. If, of course, the dictation scores were heavily biased in favour of spelling and punctuation, this would not be too surprising, but it is noteworthy that the dictation as used in this study does not relate closely to any other listening comprehension measure. (Darnell later found that clozentropy related most closely with the Listening Comprehension section of the TOEFL. This was taken by Oller to be further evidence of the similarity between cloze and dictation.)

Oller and Conrad (1971) looked at the relationship between cloze and the UCLA ESLPE Form 2C, and discovered that cloze correlated most highly with dictation. Their figures are as follows:

	Vocabulary	Reading	Grammar	Article	Cloze
Dictation	.59	.80	.60	.17	.82
Cloze	.59	.80	.58	.33	

The sample size was 35 only. Unfortunately, again no details are given of the nature of the dictation, or of how it was scored. Nor are any details given of the relative weightings of the parts of the ESLPE. As we have already seen, and as we will see again, this is most important. It is nevertheless interesting that cloze and dictation seem to have approximately the same correlations with the other subsections of the battery, although, of course, they could be measuring different parts of the variance of these subsections and they are both closely associated with the reading test.

Oller (1972) is primarily concerned with the cloze procedure, and so does not report the intercorrelations of dictation and the rest

of the UCLA ESLPE Form 2A Revised used in the study. He does, however, report the correlations between cloze and dictation for three different difficulties of text, and two scoring methods (exact and acceptable) of the cloze. No details are given on the nature of the dictation passages or the scoring procedure. The correlations range (there are eight in all) from .68 to .85. The higher correlations are for more difficult texts and the acceptable scoring method. In fact, in every case, using the acceptable method to score cloze increases the correlations with dictation. Moreover, the cloze correlations with dictation are higher than with any other subtest; only the Total, as expected, correlates more highly with cloze. However, the weighting of the subtests is known this time: Vocabulary, Reading and Grammar all weight 40 points; dictation weighs 90 points. One would thus expect any correlation of dictation and the Total to be fairly high, and similarly, cloze correlations with dictation and the Total would be expected to be similar. One would also expect higher correlations with a test which spreads subjects out over a wide range (i.e., 90 points), than one which has a lower range and, therefore, spread (i.e., 40 points). This could be the reason for the way the three 40-item tests correlate with cloze at approximately the same level, and that dictation correlates at a higher level. One must conclude from the evidence reported that dictation and cloze are related, and perhaps more so than the other subtests, but that we cannot be sure. To have added more to our knowledge about dictation, this study would have had to report the correlations with the other subtests, to have given us more information about the dictation, and to have allowed for differential weighting of the subtests. Oller also ran partial corre-

lations of the cloze tests, and the four sections of the ESLPE, and in every case dictation had the highest correlation with cloze. Conversely, when dictation was held constant, the other three subsections had lower correlations with cloze than when other subsections were held constant. All of which implies a reasonably close relationship between cloze and dictation, if it were not for the caveat mentioned above. Oller says that this close relationship is due to the fact that dictation, like cloze, and unlike the other subsections of ESLPE, is a complex, integrative task which requires "active hypothesis testing and analysis by synthesis . . . Both the cloze tests and the dictation require analysis by synthesis where the examinee has to generate responses," since the testee's ability to do dictation is "very much limited by his own capacity to rapidly synthesize meaningful sequences in the language."

Before looking more closely at what Oller claims dictation is testing, one should finally report on a study carried out by Oller, Atai and Irvine (1974) which looked at cloze, dictation and TOEFL. This time there are details of the passages used - one thought easy, one difficult; the procedure used - the dictations were read three times in all; and the scoring procedure is given - the number of words appearing in the original sequence were counted. Misspelled words were not considered incorrect as long as no phonological rule of English was being violated - e.g., "comunity" was considered correct, but "proplem" for "problem" was not. Presumably this dictation cannot be said to be a test of spelling. The question of weighting mistakes is not raised. Briefly, the results showed a fairly high intercorrelation of the two dictations (.85) (given that the two dictation passages had a high intercorrelation, and showed

little difference in their correlation with the TOEFL subtests, only the total dictation correlations will be reported) as well as fairly high correlations for the dictation total with both exact word cloze (.69) and acceptable cloze (.75). Of the TOEFL subtests, dictation correlated most highly with the Listening Comprehension subtest, and lowest with Vocabulary (.69 and .47 respectively). However, it also correlated moderately highly with English Structure (.63). Reading Comprehension and Writing Ability were more closely related to cloze (.67 and .66 respectively) than to dictation (.53 and .52). For the cloze, the highest correlation was with the Listening Comprehension subtest (.76), higher even than the correlation with dictation. The researchers conclude that TOEFL provides little information other than that provided by cloze, the dictation and the Listening subtest, therefore one should not separate skills and components of skills in a test battery; one should use, instead,"task-oriented tests that require the pragmatic use of language for communicative purposes". Whatever one thinks of that, it is reasonably clear that dictation is closely related to various measures of EFL ability, one of which is the cloze. Why this should be is perhaps less clear.

In the 1971 article Oller maintains that dictation as a valid testing technique is supported by theory as well as data. He is referring to theories of listening comprehension which assert that speech perception is an active process. The listener must extract an intended set of words from a sequence of sounds during dictation, and in effect he reconstructs the message. This is why, he claims, students reproduce "scientists from many nations" as "scientists' examinations", or "they never made" as "they are never made".

Clearly the student is not recording information faithfully, like some automatic mechanism, but is in some fashion processing the sounds he hears into words, phrases and sentences, and, presumably, interpreting as he proceeds. Oller suggests that a dynamic process of analysis by synthesis is involved. In other words, the processes required for successful performance on dictation are the same as those required for listening comprehension, and so dictation is justified theoretically. (Of course, one has no certain knowledge that analysis by synthesis is required for either dictation or listening comprehension.) Oller does not consider the next stage, involved in dictation but not in listening comprehension, namely, the transfer from sound to marks on the page. He merely claims that "dictation tests a broad range of integrative skills", without specifiying what these are. He goes on to associate analytical objective tests with Bloomfieldian/Chomskyan linguistics, and maintains that these views treat language as a self-contained unit, apart from communication.

> "If it is indeed true that language cannot be successfully explained apart from its use as a medium of communication, it would follow that analytical tests of language competence which remove linguistic units from the meaningful contexts in which they occur are apt to be less valid than integrative tests which are more relevant to communication skills." (i.e., dictation)

Why this should rule out more traditional listening comprehension tests, such as ELBA Part 1, Test 4, is not clear. "Dictation is apt to provide a more comprehensive sampling of the integrative skills involved in the understanding of complex English structures than the more isolative and

analytical objective tests." Yet of the skills Oller says the student is tested on in dictation, namely, 1) the ability to discriminate phonological units, 2) make decisions concerning word boundaries in order to discover sequence of words and phrases that make sense (i.e., that are grammatical and meaningful), and 3) translate this analysis into a graphemic representation, only the last one is required by dictation alone, since the others are clearly involved in any test of listening that uses sentences. Yet only the last skill (the ability to translate this analysis into a graphemic representation) is unaccounted for in his theoretical account of what happens in a dictation.

However, Oller dismisses the criticism that we do not know what dictation tests by saying,

> "If we knew all the psychophysical details of the process, we would no doubt soon have a listening machine for the deaf and a reading machine for the blind . . . Is it necessary to know exactly <u>what</u> a test is a test <u>of</u> in order to make use of it?"

In a later article (Oller and Streiff, 1975), he develops his theory of the nature of dictation somewhat, and quotes Neisser (1967), Cooper (1972), Liberman et al (1967) and Chomsky and Halle (1968) as proposing and supporting a model of active, analysis-by-synthesis speech processing. He then goes further, and claims that the listener's mechanism for comparing the synthesis with the incoming sequence of sounds is a grammar of expectancy. "The perceiver formulates expectancies (or hypotheses) concerning the sound stream based on his internalised grammar of the language" - i.e., dictation is a measure of the efficiency of grammar-based expectancies.

In this model, the listener/dictatee forms a crude notion of what is being talked about, and then analyses "in a deliberate, attentive, sequential fashion" in order to put on paper what he has heard. Examples like "scientists from many nations" becoming "scientists' examinations" are proof of an active analysis by synthesis.

"Since the dictation activates the learner's internalized grammar of expectancy, it is not surprising that a dictation test yields substantial information concerning his overall proficiency in the language."

This, of course, leads to the conclusion in Oller, Atai, and Irvine (1974) that

> "the test modality has a negligible effect on the results when what is being measured taps a source necessarily common to both writing and speaking skills, namely, the learner's underlying language competence, or internalised expectancy grammar."

Hence the high correlation between dictation and cloze.

Have the experts been refuted? Is everything said by Lado, Harris, Rivers and others no longer valid? The question of dictation's reliability has not been dealt with, and thus we cannot conclude that dictation is or is not reliable. The primary concern has been with validity.

Rivers' claim that dictation is redundant is still true, since dictation appears to correlate highly with other types of tests. It would appear, therefore, to add nothing to the information already being provided by various test batteries as to individual language proficiency. Nor has Oller countered the claim that dictation is un-

economical. Although he maintains in Oller and Streiff (1975) that there is no "dead data" in a dictation, he produces no evidence to prove this. It is likely that, although it is true that a student may make a mistake anywhere in the text, most students do not, and for any one student, most of what he writes is correct. Harris' claim that dictation is imprecise would be claimed as a virtue by Oller if by imprecise is meant "not discrete". Similarly, that dictation is not a valid test of listening comprehension alone, would be regarded by Oller as an advantage, and as further proof that dictation is an integrative test. Lado's point that dictation does not test aural perception is in any case doubtful, since part of what is called listening comprehension is the ability to use context in order to recover misheard words and one cannot use the context without understanding it. Similarly, his point that the order of words is given by the tester is true for the speaker, as Oller points out, but not necessarily for the listener, who must in some sense reconstruct the message. Mistakes made on dictation by students often involve changes of word order. In general, Lado's claim that dictation measures very little must now be disputed.

Apart from Lado's points referred to above, what was said by other experts remains substantially valid. What has happened to change things is twofold. First, there has been a change in testing values and objectives. It is no longer unambiguously important to isolate a skill or subskill for testing purposes, especially since it has become clearer that this is rarely possible. Spolsky (Jones and Spolsky, 1975) assumes the battle between discrete-point and integrative tests has almost been won: "With most of the big guns now on their side, the

integraters have not yet squelched some discrete practitioners." Thus what was said by previous experts, the discrete practitioners, about dictation is still true, but its implications are different. Secondly, some empirical data on dictation has at last become available. Previously, the subject had been dismissed out of hand, largely because of the prevailing intellectual climate. Even Valette, whose results are now quoted as proof of the value of dictation, hesitated to recommend dictation as a testing technique.

And what, in conclusion, does this data tell us about dictation?

1) It seems clear that dictation is related to overall language proficiency, since correlations of from .78 to .94 have been achieved with tests of linguistic proficiency.
2) Dictation correlates highly with composition (.72), vocabulary (.65), a reading survey (.78), cloze (.82, .86, etc.), clozentropy (.63), reading (.80), other dictations (.85), listening comprehension (.76) and English structure (.63), at least.
3) It correlates less highly, and in some cases poorly, with phonology (.57), oral comprehension (.59), oral interview (.48), article (.17), vocabulary (.61, .47) and grammar (.6).
4) The correlations with cloze range from .68 through .75 to .85.

The evidence from 2) and 3) is contradictory and thus inconclusive. Dictation appears to correlate both high and low with vocabulary, structure, and listening comprehension. It correlates both high and low with discrete tests (vocabulary, grammar, possibly even listening comprehension) and integrative tests (composition, oral

interview, probably oral comprehension). The only consistent thing to be shown is a high correlation with reading tests and with cloze. This hardly helps us to discover what the dictation is testing. It is not enough to claim that the dictation is associated with reading ability, since even phoneme discrimination tests usually require reading ability, as well as the ability to pronounce correctly what one reads. It is debatable whether so-called discrete-point tests really test discrete skills, in which case to claim that dictation tests integrative skills is meaningless. In reply to Oller's claim that integrative tests use language in context, two points must be made. First, Rivers' point that students taking dictation do not pay attention to either the meaning of what they are writing or, more importantly, to the way the segments fit into the whole passage. This merits further study. Secondly, it is rarely true that analytical tests remove language from its context. What, in any case, is context? Is it quality or quantity of surrounding and constraining matter? Is a sentence not adequate? Are 150 words of context really enough? As for Oller's theory of listening comprehension and the way dictation works, again, two points need to be made. First, the theory does not explain why some tests of listening comprehension do <u>not</u> correlate highly with the dictation. It is not good enough to conclude therefore that they cannot be tests of listening comprehension. Secondly, do we really need a grammar of expectancy to explain either dictation, or its relationship with language proficiency measures or with cloze?

Finally, what this review has shown, if nothing else, is that further investigation of dictation is needed, and that when

reporting on dictation one must indicate exactly how the tests were scored. Moreover, an indication of the relative frequencies of phonological, grammatical, lexical and orthographic mistakes might give some idea of what the test is actually testing, i.e., what the scores consist of.

5.6.2.2 The use of dictation in the main study

In view of the foregoing it was decided to investigate the relationship between cloze and dictation and other measures of language proficiency. This would involve giving a cloze test, a measure of EFL proficiency (in this case the ELBA), and a dictation to the same students. Certain administrative and procedural problems were encountered.

First, what sort of text should be used for the dictation? Would any text do, as Oller suggests in the discussion following the 1975 paper? He claims to have taken three passages, one obviously easy, one obviously difficult, and one medium, and to have found that all three correlated similarly with "external validating criteria". Unfortunately, he gives no further details. However, in answer to a suggestion that too easy a text will result in a lot of "dead data" because students will not make any mistakes, Oller replies that even advanced students make errors in simple dictations. It was thus decided to have two texts in this study, one easy, the other more difficult. The easy text was taken from one of the passages used on the Algerian study, which had a Fog Index rating of 7 (easy), and which had proved fifth easiest on the cloze (see Appendix C). The difficult text was taken from Fountain (1974), Test C. This text was specially written by Fountain. Each paragraph was written so that it contained a specific

number of key words, taken from a specific level of frequency on the Thorndike and Lorge 30,000-word list. Each paragraph, out of a total of five, was designed to be more difficult than the preceding one. The chunks into which the paragraphs were divided for dictation also increased in length through the passage, but were equivalent within each paragraph. The syntax of each passage was made more difficult by introducing longer, more complex and less common forms of sentence construction in the later paragraphs. Finally, when the dictation was recorded, an attempt was made to keep reading speed constant for each paragraph, but to increase it for each paragraph level. Thus the difficulty of the dictation was increased in successive paragraphs by attempts to control four variables: the frequency level of the key words, the average length of the dictated chunks, the complexity of the syntax and the speed of reading. The results Fountain got indicate that there was indeed an increase in difficulty throughout the passage.

Secondly, how should the dictation be administered? The literature has two opinions, and practice in class varies; therefore, it was decided to carry out both methods. The first consisted of three stages: reading the text normally first, then chunked for dictation, then read as a whole at the end to allow revision. Pauses were made long enough to accommodate even the slowest writer. In the second method the dictation was read once only, chunked into suitable lengths. It was never heard as a passage, and no opportunity was allowed for revision. The first method was used with the easy text, and the second with the hard text. The chunking for the first method was tested with intermediate students at Stevenson College, and adjustments in length

were made where students either complained or appeared to be having difficulty in remembering because of the length.

Finally, how should the dictation be scored? Enough has already been said to indicate that the scoring system is crucial to what dictation is a test of. There are probably an infinite number of ways of scoring dictation; however, the basic unit of scoring is usually the word. Valette maintains that "only one error per word should be counted, for the student who omits a word should not be penalized less than the one who tries to write the word and makes several mistakes." She herself gives four systems, one simple: "1 point off for each incorrect or omitted word", and several more complicated, with weightings according to the "gravity" of the error. Thus her fourth method is "1/4 point off for a wrong or omitted accent, 1/2 point off for a misspelled but recognizable word, 1 point off for each omitted or unrecognizable word, 1 point off for a word containing a morphological error, such as an incorrect verb or adjective ending." Oller's two scoring systems have already been mentioned, namely counting errors in spelling, phonology, grammar or choice of word, with equal weighting, and not counting misspelled words unless they violated some phonological rule of English (otherwise all errors were considered equivalent). The system used by Fountain was to mark only the key words in the passage, ignoring therefore all grammatical words and syntactic errors, as well as those words which were not on the appropriate Thorndike-Lorge level. Each key word was given one mark if correctly spelt. Mistakes involving omission or addition of final -s, -es, -d or -ed when these forms were suffixes were ignored. This rule

was applied regardless of whether the word was possible or not (e.g. "musics"), except for irregular verbs and nouns ("womans" and "breaked" were counted as errors).

The problem of weighting errors is a serious one, but the solution must be a question of judgment. This study will adopt the following procedure:

Text 1 (Easy): Basically Oller's second system, namely, spelling errors will not be counted unless they violate some phonological rule of English. Otherwise, errors of phonology, grammar, morphology, semantics and omission will all be counted as one point deducted. A maximum of one point deducted per word. Punctuation not marked.

Text 2 (Hard): Two scoring systems. One as above, the other similar to the Fountain system, but ignoring misspellings that do not violate some phonological rule of English.

The two dictation passages were recorded on tape, and played to subjects through an extension loudspeaker for maximum clarity. Subjects were handed a sheet of paper on which they wrote their names. They were told to expect a dictation test, and that all instructions were on tape. They then heard the easy dictation, read, as described, three times in all, followed by the hard text, read once only. They were then given the cloze test to complete. The whole session took a maximum of one hour.

CHAPTER 6

Results 1) : Native Speakers

6.1 Subjects

The 360 native speakers used as subjects in this first part of the study were all Scottish school children aged approximately 15, coming from the fourth grade of five Edinburgh secondary schools, namely, Liberton High, the Royal High School, St. Thomas of Aquin's (Roman Catholic)High, Holy Rood (Roman Catholic) High, and Broughton High. The schools can be seen as providing a sample of all social classes, income groups, and abilities within the city of Edinburgh, although this is not important for the study, whose aim was simply to test reasonably competent readers. It was emphasized by the schools that the children tested, whilst possessing a range of academic abilities, did not include non- or poor readers. Sixty percent of the children were girls. Most of the children were tested in the early afternoon, at some point in the seven days from June 18 to 25, 1975. As the end of term was approaching and exams had already been taken, the children were receptive to the tests and did not resent them. The atmosphere throughout the testing sessions was friendly and co-operative. The tests were given during a normal school period, under normal testing conditions, but no time limit was set. Obviously some children finished before others, but they were discouraged from disturbing or pressuring those who had not finished. The session was always completed within a period. No student took longer than 30 minutes to complete his or her test paper.

6.2 Scoring

360 scripts were used in the analysis, 30 subjects taking one test at one deletion rate (for example, the cloze test on the difficult text, with every sixth word deleted, was done by 30 subjects). Fifty answers on each script were punched onto card, so that from these cards it was possible to produce summary tables by computer of all the different answers to each question for each test. This summary was then used as the basis for selecting the correct answers for a scoring key. Five scoring keys were produced in this manner, and computer programs were written to score the raw data files using the scoring keys. The scoring procedures used were 1) the exact word only, 2) any semantically acceptable word (SEMAC), 3) any grammatically correct word (GRCO), 4) any word from the same form class as the deleted word (IDFC), and 5) any word fulfilling the same grammatical function as the deleted word (ACFC). When the scoring was complete, it was also possible to produce an item analysis for each test scored by any or all of the five procedures.

6.3 Results

Descriptive statistics of the results of the cloze tests are provided in Table 6.00. Reference to and detailed analysis of the results will be made in the subsequent sections. Before making a detailed analysis of the effect of the various variables introduced in this study, a two-way analysis of variance was carried out on the results to check that significant effects had been achieved. If no effect was found, there would be no point in further analysis. The two independent variables were text and deletion rate, and the results are presented in

Table 6.1 for each scoring procedure.

For the exact word method, the semantically acceptable method (SEMAC), and the same grammatical function procedure (ACFC), no significant effect of deletion rate was found. The other two procedures both showed a significant effect of varying the deletion rate. A highly significant effect of varying textual difficulty was obtained, which establishes that for these texts, regardless of deletion rate, cloze is sensitive to changes in text difficulty. If one groups together the texts and looks at deletion rate, which is what the two-way analysis does, then it appears that cloze is not sensitive to changes in deletion rate. However, the validity of this procedure is doubtful because of the significant differences between texts. Since highly significant interaction effects were revealed by the two-way analysis, it is clear that at least the combination of certain texts with certain deletion rates changes the cloze score significantly. Interestingly, however, only the exact word method showed this significant interaction effect, the other scoring procedures producing effects which were either only just significant, or not significant at all. Nevertheless, regardless of scoring procedure, the F for text differences was always significant.

Despite the mixed results from this preliminary investigation, it was felt that the two-way analysis justified further examination of the results, so the effect of the three main variables - deletion rate, textual difficulty, and scoring procedures - was looked at separately for each variable.

6.3.1 Text

The null form of Hypothesis 2 (Chapter 5, section 5.1)

states that there is no difference in the ranking of texts by the cloze procedure using different deletion rates, and scoring by different procedures. Table 6.2 gives the rankings of the three tests - easy, medium and difficult - for each deletion rate, and the five scoring procedures.

T A B L E 6 . 2

Ranking of texts, by scoring procedure and deletion rate. Native speakers.

Deletion rate 6

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	34.3 (1)	46 (1)	48.4 (1)	45.6 (1)	46.7 (1)
Medium	25.5 (2)	38.5 (2)	45.6 (2)	43.3 (2)	44.8 (2)
Difficult	19.6 (3)	33 (3)	43.9 (3)	35 (3)	37.4 (3)

Deletion rate 8

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	34.9 (1)	45.3 (1)	48.4 (1)	45.4 (1)	47.1 (1)
Medium	24.9 (2)	39.5 (2)	46.6 (2)	41.3 (2)	44.5 (2)
Difficult	15.9 (3)	34.5 (3)	44.4 (3)	36.5 (3)	38.9 (3)

Deletion rate 10

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	32.6 (1)	43.5 (1)	47.2 (1)	46.5 (1)	46.6 (1)
Medium	29.8 (2)	41.7 (2)	46.5 (2)	44.9 (2)	46.3 (2)
Difficult	14.7 (3)	31.8 (3)	41.2 (3)	36.9 (3)	38.8 (3)

Deletion rate 12

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	30.1 (1)	43.4 (1)	47.5 (1)	44.6 (1)	46.5 (1)
Medium	29.0 (2)	39.1 (2)	44.1 (2)	43.1 (2)	44.0 (2)
Difficult	20.3 (3)	34.9 (3)	43.3 (3)	38.2 (3)	39.6 (3)

From this, it is clear that the texts are always ranked in the same order by the cloze procedure, regardless of any change in the deletion rate or the scoring procedure. Even scoring procedures which one would expect to be insensitive to the differences for native speakers prove to be capable of distinguishing among the three texts consistently. It is more difficult for native speakers to supply a grammatically correct word in a cloze gap in a difficult text than it is to supply a grammatically correct word in a medium text. Indeed, it is easier to supply words fulfilling the grammatical function in an easy text than in a medium text. This result is tempered by the fact that the three texts were deliberately chosen to be as different from each other as possible, in the expectation that certain versions of the cloze would not prove sensitive to their differences. Had this been the case, one could have generalised to conclude that certain versions of the cloze would prove incapable of distinguishing relative text difficulty for texts more closely related in difficulty levels. This has proved not to be the case; whether, however, cloze would be capable of distinguishing reliably among less extreme texts remains unanswered.

Although the hypothesis refers only to ranks of texts, it was decided to see whether the differences between texts were real differences, or whether they could have occurred by chance alone. For this

purpose, any difference between deletion rates was ignored, and the results for any one text summed over all deletion rates were then averaged. A one-way analysis of variance was performed on the means of the three texts, scored by all five methods, and the results tabulated in Table 6.3. From this it is clear that the texts are always significantly different from each other, regardless of the scoring procedure used.

In summary, then, cloze seems to be sensitive to differences between texts regardless of the scoring procedure used or of the frequency of deletion of words. This suggests that, at least for native speakers, cloze is a suitable measure of readability, and that the use of a different deletion rate should not produce a different rank for a text. (However, although the differences between texts that cloze produces are indeed real differences, there is no guarantee that had other, less different texts been used, the same results would have been achieved).

Regarding cloze as a test of reading comprehension, little can be concluded from this study as to the suitability of difficult rather than medium or easy texts. Since no independent measure of the reading ability of these native speaker subjects was available, it is impossible to compare the cloze with anything else. It is possible to compare different cloze versions, using different deletion rates and different texts, as tests in terms of efficiency and item effectiveness, but this will be postponed until the section on cloze as a test for native speakers (section 6.3.4).

6.3.2 Scoring procedures

The null form of Hypothesis 3a is that there is no significant difference between exact and other scoring methods. The expectation

is, of course, that different scoring methods for cloze tests will result in different mean scores, since one assumes that different scoring procedures measure different aspects of whatever the cloze procedure is a test of.

To investigate this hypothesis, paired t-tests were run on the mean of each test (i.e., each text at four deletion rates) when scored by the five different procedures. The results are summarised in Table 6.4 a, b and c.

The general result is as expected, namely, that the different scoring procedures result in significantly different scores. This is especially true for the difficult text, where all possible comparisons show significant differences.

For the easy text, this is not so, since the form class scores (IDFC and ACFC) tend not to produce scores different from those produced by the semantically acceptable procedure. At both deletion rate 6 and deletion rate 8, the semantically acceptable and identical form class scores are not different from each other; further, at deletion rate 6, the same-grammatical-function procedure does not produce scores significantly different from the semantically acceptable scores, and at deletion rate 10, the same-grammatical-function procedure results in scores which are essentially the same as the grammatically correct and identical form class scores. However, this would appear to be explained by the fact that, for the easy text, virtually maximum scores were achieved by four of the five scoring procedures. Although the medium text gave two non-significant comparisons - between the same-grammatical-function method and the any-grammatically-correct procedure

on both deletion rate 10 and deletion rate 12 - here again one is dealing with high scores (means of up to 93%).

It seems that one can safely conclude that in practice different scoring procedures give different results. These procedures seem to measure different aspects of the reading process.

It is interesting to note that even with native speakers, different grammatical scoring procedures produce different results, and this, even on easy texts, over eighty percent of the time. This seems to imply that no one grammatical scoring procedure adequately taps the native speaker's ability to respond to the syntax of a text.

However, whilst it may be true that cloze scored by one procedure is a different test from cloze scored by another procedure, this presumably only has serious practical consequences if the rank order of subjects changes. Regardless of whether different procedures result in different tests, do the subjects retain the same position relative to each other? If not, then the information provided by a different scoring procedure is effectively redundant.

The rank order of subjects on different scoring procedures was checked by the Spearman rho correlation coefficient, and the results are tabulated in Table 6.5.

The first point to be made is that the rank orders clearly do differ. Although the correlation coefficients show a great deal of variety (from .41 to .99), of the 120 coefficients only 28 are of the order of .90 or higher. This is not quite what findings of people like Oller would lead one to expect - although his studies were with non-native speakers. Other authorities claim to have discovered that

different cloze scoring procedures are closely interrelated, and that therefore the exact word procedure is preferable because it is easier to apply.

The highest amount of agreement on the rank order of subjects is achieved by the difficult text, whereas those texts which were relatively easy for the native speakers result in lower coefficients. This, of course, is partly due to the fact that if scores are closely bunched together at the top of the distribution, as tends to be the case for the easy text, than a change of even one point can result in a major change in rank order. One would thus expect easier texts to produce lower coefficients.

The lowest agreement, regardless of text, is between the exact word score and the grammatical scoring procedures. In fact, there seems to be relatively little agreement between the exact word method and the procedure scoring any grammatically correct word as correct. If the exact word method really is a measure of reading comprehension, then the ability to fill in gaps with grammatically correct words is not related to comprehension for native speakers. The exact word method is, however, much more closely related to the ability to restore deletions with words which are semantically acceptable in a particular context.

The highest amount of agreement, perhaps not surprisingly, is among the different grammatical scoring procedures. In particular, the two form class procedures (IDFC and ACFC) are consistently closely related (.90 to .99). It is fairly obvious that they measure the same thing, and one is redundant. These, however, are the only consistently close relationships.

Although, on the whole, the exact word method and the semantically acceptable method are reasonably closely related (overall, regardless of deletion rate, between .71 and .83), there are quite wide variations in the relationships, depending on the text and deletion rate being used. Thus, although one could conclude, as many have, that exact and semantically acceptable procedures are closely related to each other, this close relationship is in many cases more apparent than real, and in fact, the correlations achieved may vary quite widely depending upon the specific test. This suggests that one cannot regard the two procedures as equivalent, and that scores on one procedure are not necessarily adequately predicted by scores on the other. In other words, cloze exact and cloze semantically acceptable scores are not interchangeable and mutually substitutable. One procedure clearly provides different information from that provided by the other.

In summary, then, the following points have emerged from Table 6.5:

1) different scoring procedures do not measure the same thing;
2) grammatical scoring procedures are closely related, and at least one is superfluous;
3) the grammatically correct and exact word procedures show relatively little relationship;
4) the semantically acceptable and grammatically correct procedures show a reasonably close relationship; and
5) the exact word and semantically acceptable procedures are related, but not enough to make one of the procedures superfluous.

The final point to be considered in this survey of the

differences and similarities among cloze scoring procedures is the rank order of the procedures themselves. An examination of Table 6.00 (descriptive statistics) reveals that in eleven out of the twelve tests, the rank order of scoring procedures is (easiest first): grammatically correct; same grammatical function; identical form class; semantically acceptable; exact word (GRCO, ACFC, IDFC, SEMAC, EX). In the one case, Test E06, where the SEMAC and IDFC procedures changed positions, the t-tests of Table 6.4 showed no significant differences between the means of these procedures, so that the rank order could just as validly be reversed. Thus, effectively, all twelve tests agree that the grammatically correct method is the easiest scoring procedure for native speakers, followed by ACFC and IDFC, whilst the exact word is always the most difficult, followed by the semantically acceptable procedure. The implication of this for the grammatical scoring procedures is simply that although a native speaker may not always provide an answer from the same form class as the deletion, or one which performs the same grammatical function, the answer he does provide will tend to be grammatically correct.

The general conclusion seems to be that as the criteria for correctness become progressively narrower, the difficulty of achieving correct replacement increases, or, as the similarity of the replacement to the deletion increases, it becomes increasingly difficult to supply. As the discourse constraints on the replacement imposed by the scoring procedure become tighter, the difficulty of correct replacement increases. This, however, is not necessarily related to the <u>amount</u> of context constraining the replacement, such that a grammatically

correct word can be supplied from only two words of context, whereas to supply the identical form class needs four words of context, and to supply the exact word requires reference to the whole paragraph or discourse, or to knowledge of the world. If difficulty were related to the amount of context constraint, one would expect one scoring procedure to respond differently to changes in deletion rate from another scoring procedure. This is the subject of investigation of the next section.

6.3.3 Deletion rates

The deletion rate had been systematically varied for each text, removing every 6^{th}, every 8^{th}, every 10^{th} and every 12^{th} word respectively to produce four different deletion rates.

One-way analyses of variance were performed on the results of each text scored by the five different procedures, with deletion rate as the independent variable. The results are set out in Table 6.6a - e.

Highly significant differences between deletion rates were found for all three texts scored by the exact word method.

The semantically acceptable procedure and the identical form class method showed significant ($p < .05$) differences between deletion rates only for one of the three texts - the easy text in the case of the semantically acceptable procedure, and the medium text in the case of the IDFC.

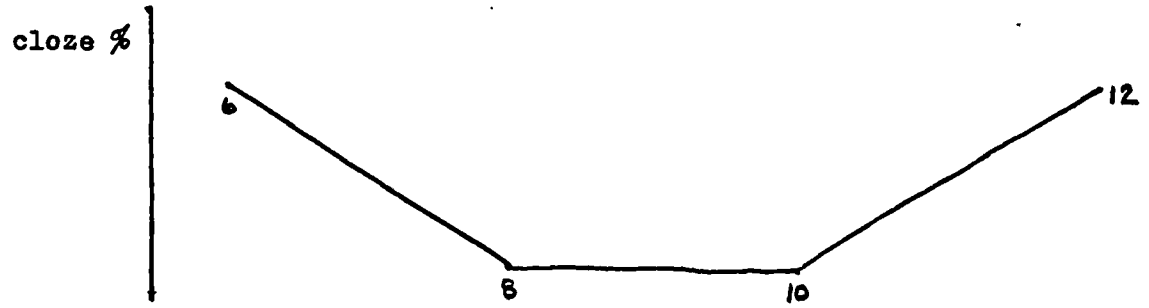
The remaining two scoring procedures (grammatically correct and same grammatical function) showed no differences between deletion rates on any text.

Having determined that there are significant differences

between deletion rates for some scoring procedures and some texts, it was then necessary to see exactly where the differences lay, since all the analysis of variance tells one is that differences do exist. The expectation was that difficulty would decrease as the deletion rate decreased, so that 12 would be easier than 10 would be easier than 8 and so on. However, an inspection of the means in Table 6.6 shows that this is rarely the case. Only once in fifteen times do the means increase regularly from deletion rate 6 to deletion rate 12 (IDFC, difficult text). More often, the means actually decrease, - i.e., the texts become more, not less, difficult as the length of context surrounding the gaps increases (Figure 6.1). The one-way analyses of variance showed, however, that ten of these fifteen graphs can be discarded, since no statistically significant differences were found.

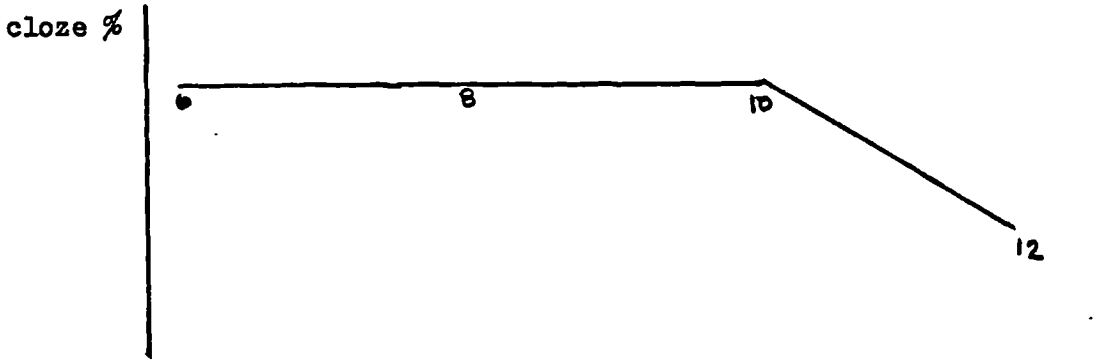
In order to see, from the remaining five tests, which deletion rates were significantly different from which others, t-tests were run for all possible pairs. The results are tabulated in Table 6.7.

To take the exact word method first, on the difficult text all possible pairings were different, except for 8 and 10, and 6 and 12. Deletion rate 6 was different from 8 and 10, and deletion rate 12 was different from 8 and 10. This can be graphically represented as:



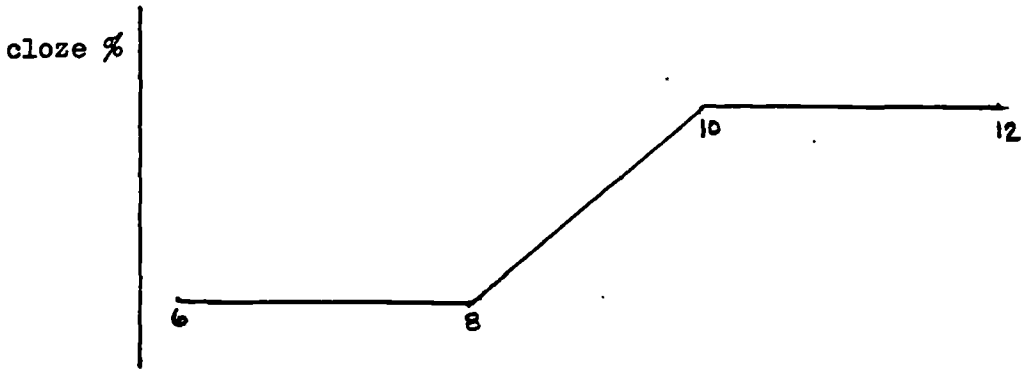
where positions on the same level show no significant differences. The middle two deletion rates are significantly harder than the two extreme deletion rates.

On the easy text, however, only deletion rate 12 was significantly different from the other three. Graphically, this is shown as:



Put into words, the test with the longest context for each gap was the most difficult, contrary to expectations.

The medium text produced yet another picture, but one closer to the expectations, since 6 and 8 formed one group, and 10 and 12 a second, and the members of each group were significantly different from both members of the other.



Here, at least, as deletion rate increased, the difficulty of the cloze test decreased.

For the exact word method, then, no consistent pattern has

emerged other than the already established fact that deletion rates do differ. Deletion rate 6 is sometimes different from 8, 10 and 12, and sometimes it is not; sometimes deletion rate 10 is the same as 12 or 6 or 8, and sometimes it is different. The only consistent fact is that deletion rate 8 is different from deletion rate 12, and on that no theory should be built, since in one case it is easier than 12, but in the other two cases it is more difficult.

Table 6.7b shows the results of the t-tests on pairs for the semantically acceptable score, easy text, and the medium text scored by the IDFC procedure.

For the former, deletion rate 6 is different from 10 and 12; otherwise all pairs show no significant differences. Thus there is some increase in difficulty as the deletion rates increase, but not very much, and in any case, the means are so high (87% to 92%) that this effect is probably negligible.

For the identical form class scoring procedure, deletion rate 8 is different from 6 and 10, but otherwise all pairings are the same - i.e., the only increase in difficulty occurs at deletion rate 8. Again, the means are high (83% to 90%), so the foregoing caveat also applies here.

The conclusion thus far seems to be that significant differences do exist between deletion rates, but that the differences are neither consistent nor predictable. However, using any scoring procedure other than the strictest (viz., the exact word) drastically reduces, and indeed usually removes, differences between tests due to deletion rates.

However, as was pointed out earlier (Chapter 4, section 4.6.1), the difference between tests at different deletion rates is not purely a difference of length of context between gaps. Inevitably, to maintain the same number of items, a deletion rate of 12 has twice as much text as a deletion rate of 6, so the texts are appreciably different. Also, the deletions are not the same throughout, since different words are of necessity deleted by different deletion rates. It is, however, possible to take only those words deleted in both tests of the pair one is considering, and then to compare the means based on those items alone. Thus, since counting for deletions always started at the same point, item 2 in deletion rate 6 is the same as item 1 at deletion rate 12, and in the comparison 6:12, 25 items are common to both tests. In the comparison 8:12, there are 16 items in common; in 10:12, 8 items in common; and so on.

Computer programs were written to select only those items common to both pairs of any comparison, calculate the means and deviations based on those selected items, and make t-tests for differences between the means. These calculations were done for the exact word score for all texts; the semantically acceptable score, easy text; and the identical form class score, medium text - i.e., those tests where the analysis of variance had shown significant differences between deletion rates. It was assumed that for the other tests, differences between deletion rates, even for identical items, did not exist. The results are presented in Table 6.8.

From these results, it is immediately apparent that if non-identical items are excluded from the tests, no differences in deletion

rates are to be found. Moreover, this is true whether one scores by the exact word method, the semantically acceptable method or the identical form class method. Admittedly, in the case of the latter, three of the six comparisons show significance, but the means are so high, and the deviations so low, that one cannot place complete confidence in the results. Furthermore, some items were of the nominal modifier type - e.g., the green car - where, under this scoring procedure, a noun used legitimately as a modifier would be counted as incorrect. This tends to distort the results.

From these results, it is possible to draw the following conclusion. Increasing the amount of context on either side of a cloze gap beyond five words has no effect on the ease with which that gap will be clozed - for native speakers. No increase in predictability is gained by a bilateral context of eleven words rather than five words, and this is true not only for the subject's ability to respond with a semantically acceptable word, but even for his ability to respond with the exact word deleted. If amount of context has any effect, the critical amount is less than five words. This confirms MacGinitie's finding that increasing context beyond four words has no effect on the predictability of a word. Whether this is also true for non-native speakers remains to be seen.

6.3.4 Efficiency of cloze as a test

One aspect of this study which has not yet received consideration is the efficiency of cloze tests as proficiency tests for native speakers, and the influence of the three variables of text, deletion rate and scoring procedure on this efficiency. It might,

indeed, prove possible to recommend a "best buy" in cloze tests for the native speaker.

In view of the lack of external criteria against which to evaluate the cloze tests, it is necessary to take internal criteria, and, specifically, to do this by means of item analysis, reliability measures, and consideration of the descriptive statistics for each test.

Clearly, the definition of an efficient test depends entirely upon what sort of test it is, the use to which it will be put, and the population for which it is intended. One would expect a criterion-referenced achievement test to have a high mean and relatively little dispersion, whereas a norm-referenced proficiency test would be expected to have a much lower mean and wide dispersion. Yet again, a norm-referenced achievement test might ideally have a bi-modal distribution, with the pass/fail mark in the dip between the two peaks. In such a case, the distribution of either curve would be relatively unimportant, provided one had a clear distinction between the pass group and the fail group. It is thus, in principle, impossible to prefer a test with a mean of 30% and a narrow distribution to a test with a mean of 70% and a wider distribution, unless one knows exactly how the test is to be used.

Since it is impossible to say that the cloze tests have performed better or worse <u>as tests</u> than some other test, the remarks on efficiency that follow are inevitably tentative. This lack of precision is increased by the lack of guidelines, objective or otherwise, as to exactly what a good test for a particular purpose should look like. The assumption underlying the discussion, which may not correspond to the requirements of many cloze test users, is that the cloze tests are to be

used as some sort of norm-referenced proficiency test - proficiency in reading, i.e., a measure of reading comprehension - and the ability to understand written text is assumed to vary quite widely within the population. This must be qualified, however, with regard to the different scoring procedures, since the grammatical scoring procedures (ACFC, IDFC and GRCO) are not intended to provide good measures of overall reading comprehension. The interest in these procedures at this point is rather to see how using them changes the nature of the cloze as a test.

Examination of Table 6.00 reveals that for native speakers, the use of grammatical scoring procedures - GRCO, IDFC, ACFC - results in high means, although not maximum means with little dispersion. The only cases where the mean is less than 80% occur with the form class procedures on the difficult text. Nevertheless, even here, minimum scores are above 50%, suggesting relative inefficiency in the test.

The semantically acceptable scoring procedure (SEMAC) appears to be somewhat more efficient, in that its means are lower than those for the grammatical procedures, and the standard deviations are considerably larger. It would seem to be better at discriminating among subjects, and indeed the minimum scores gained on this procedure are markedly lower. Since the maximum scores attained under this procedure remain more or less the same as those gained under grammatical procedures, regardless of text, what has happened is that the distribution has actually changed shape, as well as moved lower down the scale. Mean scores, however, even on the difficult text, are rather high (difficult text, 62-70%; medium, 77-84%; easy, 87-92%), indicating the relati

inappropriacy of such a scoring procedure with native speakers.

The best distribution of scores is consistently gained by the exact word method. Table 6.00 (B) shows that, expressed as a percentage of the mean, the standard deviation is virtually always higher for this procedure than for any other scoring procedure. The means are also always considerably lower when the cloze test is scored by the exact word method. In the case of the difficult text, this results in somewhat low means (29-40%), but even here the distribution is greater than that of any other scoring procedure. It is relevant to the discussion to point out that even with an easy text and with native speakers, the cloze exact mean does not go above 70%, and the maximum score does not reach 90%.

As the discussion of deletion rates in the preceding sections has shown, there are no generalisations possible about deletion rates across texts, since the difference between deletion rates seems to be entirely due to the fact that different words have been deleted, rather than to the fact that there is a consistent difference between deletion rates. There is, therefore, little point in comparing the efficiency of cloze tests across deletion rates. Since, however, the texts are consistently different, it is possible to compare the efficiency of cloze tests across texts, holding deletion rates constant.

Bormuth (1968a) related cloze to conventional reading comprehension tests, and identified three levels of reading: the independent level - which he claimed corresponded to a reading comprehension score of 90% - which he fixed at 57% for cloze; the study level - reading comprehension score 75%, cloze score 44%; and the frustration

level - presumably below 75% on conventional tests, and below 44% on cloze. (By cloze, Bormuth means a deletion rate of every 5th word, scored by the exact word method.)

The validity of his conventional reading comprehension levels need not be considered here, nor need we consider the validity of his identification of three levels with certain cloze scores. The only point to be made is that it seems to be the case that one would expect lower means on an exact cloze test than on a traditional multiple-choice comprehension test. Unfortunately, Bormuth does not give any advice as to the ideal mean and distribution for cloze tests as reading proficiency tests, but it is possible to assume that he would identify the study level as the appropriate area, and thus cloze scores of between 44% and 57% as being the appropriate range for the mean.

If one were to take 50% as being appropriate for a reading proficiency test, then the medium text, exact score, seems to fit the bill. The difficult text scored by the exact method results in means which Bormuth would identify with the frustration level, indicating that it might be inappropriate as a proficiency test; whereas the semantically acceptable procedure on the same text results in means of between 60% and 70%. Interestingly, however, the difficult text results in a wider distribution, which might be felt to be more suitable for a proficiency test, than the medium text, which produces standard deviations varying from 13% to 19% of the mean.

It is also interesting to note that the standard error of the mean for the exact word method is greater than that of other scoring methods for all the texts, and that it increases steadily as a proportion

of the mean as the difficulty of the passages increases. This is presumably undesirable. The standard error of measurement for the exact word method is around 2.5 for all the tests, although as a proportion of the mean it increases with increasing text difficulty (Table 6.9).

The normal reliability estimate - KR21 - was unsuitable for these tests, as it is only appropriate for tests with items of similar difficulty levels. As will be seen in the next section, this is far from being the case with cloze tests. As Bormuth (1965b) pointed out, cloze item difficulty distributions tend to be U-shaped, and the tests examined here proved no exception to this. Instead, formula KR20 was used to estimate the reliability of the cloze tests, although Guilford says that "it gives an underestimate where there is wide dispersion of item difficulties" (Guilford 1965). In view of the high means for four of the scoring procedures, the coefficients were calculated only for the exact word method. The results, as displayed in Table 6.00, show a tendency for reliability to decrease as the text becomes easier. It can also be seen that different cloze forms - i.e., different deletion rates on the same text - may result in widely differing reliability coefficients (e.g., medium text, deletion rate 8: KR20 = .40; same text, deletion rate 12: KR20 = .73). In general, the reliability is somewhat low, at around .70, even bearing in mind Guilford's caveat. It is probable that higher reliabilities would be gained if the number of items in the tests were increased, but this variable is beyond the scope of this study.

Table 6.10 presents a summary of the item difficulties for all twelve tests scored by the five different methods.

If one takes the cutoff points of 20% and 80% as representing the extremes of item difficulty for efficient items - and, although these figures are arbitrary, they represent fairly normal testing practice - then it is clear that the three grammatical scoring procedures produce highly inefficient tests when applied to easy and medium texts. This, of course, is no surprise in view of the high mean scores gained by use of these procedures. Their efficiency increases on the difficult text, but rarely, even on this text, do more than 50% of the items come within these limits. The preponderant tendency is for items with a facility index exceeding 80%.

The same tendency is seen with the semantically acceptable score on easy and medium texts, so that for Tests M6 and M12, 50% of the items scored by SEMAC exceed 80% facility. On the difficult text the number of items falling within the 20% and 80% limits is greater than the number of extreme items, although there is still a marked trend towards easy items.

Only once does the exact word score result in more than 50% of items at one extreme of the difficulty scale (viz., D10). However, even this procedure gives, at best (E6), 64% of items within the acceptable limits for difficulty. In contrast to the other scoring procedures, however, the exact word method results in many items of zero facility - i.e., where no subject had supplied a correct answer. This is particularly marked on the difficult text, where, for example, on Test D10, no fewer than 38% of the items had zero facility. Even the easy text provided some items (one for E8, two for E10) where no subject was able to supply the exact word. It is apparent that all

the cloze tests and all the scoring procedures produce many inefficient items, and frequently show a large range of item difficulty.

It is possible to compare tests by examining the ratio of acceptable items to extreme items, to get and indication of which test produces more efficient items.

TABLE 6.11

Performance of scoring procedures, native speakers, in terms of item difficulty

	Best	Worst
D6	SEMAC	GRCO
D8	SEMAC	GRCO
D10	SEMAC	GRCO
D12	SEMAC	GRCO
M6	Exact	GRCO
M8	Exact	GRCO
M10	Exact	ACFC
M12	Exact	GRCO
E6	Exact	GRCO
E8	Exact	GRCO
E10	Exact	GRCO
E12	Exact	GRCO

Comparing scoring procedures across all the tests, the grammatically correct procedure emerges as the procedure which consistently produces the fewest acceptable items and the most extreme

items. (Table 6.11) From the point of view of item analysis, it appears that the GRCO is entirely inappropriate for use in a proficiency test for native speakers.

In terms of the item analysis, the scoring procedure which produces the best test is the exact word method for the easy and medium texts, but not for the difficult text. For this text, the semantically acceptable procedure gives more items of medium difficulty, and fewer extreme items, than the exact word method. The most noticeable difference between the two is that where the exact word has a large number of items of low facility, the semantically acceptable method results in a smaller number of items with high facility. The "worst buy" in scoring procedures, for a proficiency test for native speakers, would appear to be the grammatically correct method, whilst the "best buy" is the exact word method for easy and medium texts, and the semantically acceptable method for difficult texts.

T A B L E 6 . 1 2

Best performance of deletion rate, in terms of item difficulty, native speakers.

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy text	6	12	12	12	12
Medium text	12	12	12	6	12
Difficult text	8	10	10	6	6

It is also possible to compare all deletion rates for each text and each scoring procedure (Table 6.12). From this it would

appear that, at least for the easy and medium texts, the most consistently best deletion rate is the 12th. For the difficult text, the position is less clear, with a weak preference for deletion rate 6. From the point of view of item difficulty, it appears that the "best buy" in deletion rates is the least frequent deletion for easy and medium texts, and a more frequent deletion rate for a more difficult text.

TABLE 6.13

Best performance of text, in terms of item difficulty, native speakers

Deletion rate	Exact	SEMAC	GRCO	IDFC	ACFC
6	E	D	D	D	D
8	E	D	D	D	D
10	E	D	D	D	D
12	E	D	D	D	D

Finally, it is also possible to see which of the three texts produces the best distribution of item difficulties (Table 6.13). This shows quite unambiguously that for all the scoring procedures except one, regardless of deletion rate, the difficult text results in the most favourable distribution of item difficulties. However, the exact word method consistently gives its best item statistics with the easy text. The worst item distributions are a mirror image of this picture, so that for the exact word method, the difficult text is always worst, whereas all other procedures give their worst results with the easy

text. Thus the "best buy" in texts, for native speakers, is the easy text if the exact word method is being used; otherwise, regardless of scoring procedure, the difficult text.

To summarise these recommended "best buys", if the exact word method is being used, then the text should preferably be an easy one for use with native speakers. With such a text, the best deletion rate is probably the every-12th-word system. If, on the other hand, the semantically acceptable procedure is being used for scoring, then a difficult text should be chosen, with a deletion rate less frequent than every 12th word, and probably every 6th word.

The final stage in this item analysis was to calculate the item discrimination indices for the exact word procedure, using E_{1-3}. The results are set out in Table 6.14. From this it is clear that there are never more than 50% of the items in a cloze test for native speakers with a reasonable discrimination index, and there are almost always items which discriminate in the wrong direction (negative discrimination). In general, the best discrimination is achieved by the difficult text when comparing texts, and the every-6th-word deletion rate when comparing deletion rates.

When item difficulty and item discrimination are considered together, selecting only those items between 20% and 80% facility and above .2 discrimination, then the number of acceptable items is fairly small, ranging from 18% to 38% of all items. On the average, from a cloze test of 50 items scored by the exact word method, only 14 items prove to be within the conventionally acceptable limits for proficiency tests.

CHAPTER 7

Results 2) : Non-native Speakers

7.1. Procedure

The procedure for the compilation of the test booklets was exactly the same as for the native speakers. Each non-native took one cloze test, distributed in the same way as for natives - ie., in sequence, with the different texts following each other at any given deletion rate, to minimise the opportunities for cheating. The cloze test was in most cases preceded by two dictation tests, which took about half an hour. Students were allowed as long as required to complete the cloze, which was about twenty minutes on the average, the maximum being, as for the native speakers, about thirty minutes. Because of the varying times needed to complete the cloze, it was necessary that the dictation tests always precede the cloze, otherwise students would have had to wait for the others to finish before the second part of the test could be carried out. Most students took the cloze/dictation tests within a week of taking ELBA (English Language Battery), but some students had to take the tests immediately after the ELBA for administrative reasons. No complaints were noted, nor any sign of undue fatigue or discomfort.

7.2. Subjects

All subjects were aged at least 18, and were either students of English or linguistics in British or European universities, colleges of education, or colleges of further education, or they were under- or post-graduates of other subjects (medicine, engineering, anthropology,

chemistry, etc.), currently studying in the United Kingdom.

No student was coerced into taking the dictation/cloze tests, but their attendance for the ELBA was usually compulsory, being a condition of matriculation for study at the institutions concerned (Aberdeen, Bradford, Edinburgh and Newcastle universities). An attempt was made to entice pure volunteers by advertising the experiment in Edinburgh and Bradford, and by offering free refreshments in the case of Edinburgh, at the end of the ELBA testing sessions. This method produced only 40 volunteers. (For the advertisement, see Appendix E.) In all other cases students were contacted either shortly after they had taken ELBA, and asked to take part in an experimental test immediately (few refused), or during normal class hours, when the test provided welcome variety in the normal timetable. All subjects were willing to participate, and were interested in the aims and results of the study. For administrative reasons, however, some of the students taking ELBA and cloze were unable to take the dictation tests.

Three groups of students (Moray House; Stevenson/Anniesland Colleges; summer language school students) were unable to take ELBA, and thus only took the dictation and cloze tests. For technical reasons, one small subgroup of one of these groups was unable to take the dictation test, so that only cloze scores are available for them.

Those subjects who failed to complete 50% of the cloze test were rejected from the study. This left 360 subjects who had been tested on cloze, giving 30 subjects for each test. The following table gives details of the number of subjects taking each test (Table 7A).

TABLE 7A

Summary of number of subjects taking the various tests.

Taking cloze: 360

Taking cloze + ELBA: 264

Taking cloze + ELBA only: 67

Taking cloze + ELBA + dictations: 197

Taking cloze + Dictation I + Dictation II: 275

Taking cloze + Dictation II: 1

Taking cloze + dictation only: 79

Taking cloze only: 17

These 360 subjects come from the following institutions: Aberdeen University, 31; Bradford University, 9; Edinburgh University, 153; Newcastle University, 60; Stevenson College of Further Education, 22; Anniesland College of Further Education, 18; Moray House College of Education, 37; summer language courses for European university students, 30. It can thus be reasonably claimed that this group represents a fair selection of adult foreign learners of English studying in the United Kingdom. The only group deliberately excluded from the study were those learners who could be classified as beginners, elementary, or lower intermediate, since it was felt that all of the tests involved would be too difficult for them. The subjects tested could be classed, therefore, as intermediate to advanced, i.e., from just below Cambridge First

Certificate of English upwards. The mean ELBA score was 166.8, with a standard deviation of 38.6 - i.e., mean 62%, standard deviation 14% - which indicates a reasonably homogeneous, moderately proficient group.

For five of the twelve cloze test subgroups (E10, E8, M6, M8, M10) all 30 subjects have at least one measure of EFL proficiency, be it dictation, ELBA, or both. For the remaining seven subgroups, the numbers of those having at least one proficiency measure are as follows: E6, 28; E12, 28; M12, 29; D6, 26; D8, 27; D10, 28; D12, 27.

Since comparisons will be made between deletion rates for any text, and statements made regarding their ability to measure EFL proficiency (Hypothesis 4a), it is important to establish the homogeneity of the twelve groups on the measures of proficiency used. One-way analyses of variance were performed on the four deletion rate groups for each text, on the ELBA scores, the first dictation test, and the second dictation, and F ratios calculated for each of these nine analyses (Table 7B). No ratio was significant, i.e., no significant differences between the various groups on these measures of proficiency were found. It can thus be assumed that no one group is more proficient in English than another, as traditionally measured. If differences in cloze scores are found, these can be presumed to be due to the differing efficiency of cloze tests as measures of proficiency. Similarly, if some cloze tests rank the subjects in a different order from that obtained on the traditional measures of proficiency whilst others rank them in a similar manner, one can conclude that different cloze tests measure different things, or that they measure proficiency differently.

Summary

As in the native speaker study, 360 non-native speaker subjects were tested, of intermediate to advanced proficiency. This gave 30 subjects for each text, with no significant differences between groups for EFL proficiency (at least one measure of which was obtained from all but 17 subjects).

7.3 Scoring

As in the native speaker study, 360 scripts were used in the analysis, with 30 subjects taking one text at one deletion rate. The 50 answers for each individual were punched onto cards, which enabled summaries of the different responses to each item to be made by computer. From these summaries five scoring keys were produced, and the tests were scored by computer using these keys. In fact, the keys were identical in every respect, including actual words, to the keys used for scoring native speaker tests, and are labelled in the same way; viz., 1) exact word, 2) any semantically acceptable word (SEMAC), 3) any grammatically correct word (GRCO), 4) any word from the same form class as the deleted word (IDFC), and 5) any word fulfilling the same grammatical function as the deleted word (ACFC).

7.4 Results

Table 7.00 gives a summary of the results for each test, scored by all five scoring procedures. From this table, it is clear that different texts result in different means, as do different deletion rates and different scoring procedures. In order to establish that the apparent differences between texts and deletion rates were real and

statistically significant, a two-way analysis of variance was made on the results for each scoring procedure. The results of these five analyses are presented in Table 7.1a-e.

From Table 7.1 it is clear that significant differences do exist between the various tests, and in particular that, regardless of scoring method, the three texts are always significantly different from each other. On the other hand, changing the deletion rate appears to have no significant effect on cloze scores for three of the five scoring procedures, the two exceptions being the grammatically correct (GRCO) and the identical form class (IDFC) procedures. As with the native speakers, however, it is likely that the analysis of variance obscures differences between deletion rates by ignoring text differences. The interaction between text and deletion rate proved significant for all but one scoring procedure (IDFC), so it is clear that the effect of the three main variables bears further investigation.

7.4.1 Text

T A B L E 7 . 2

Ranking of texts, by scoring procedure and deletion rate. Non-native speakers.

Deletion rate 6

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	30.3 (1)	42.4 (1)	45.9 (1)	44.3 (1)	45.7 (1)
Medium	20.1 (2)	30.3 (2)	39.1 (2)	38.9 (2)	39.8 (2)
Difficult	14.8 (3)	24.9 (3)	37.8 (3)	31.2 (3)	33.0 (3)

Deletion rate 8

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	29.8 (1)	40.7 (1)	44.6 (1)	42.7 (1)	44.7 (1)
Medium	19.9 (2)	31.3 (2)	40.4 (2)	37.5 (2)	41.2 (2)
Difficult	10.0 (3)	21.6 (3)	31.3 (3)	25.6 (3)	27.8 (3)

Deletion rate 10

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	30.8 (1)	40.3 (1)	44.2 (1)	44.6 (1)	44.7 (1)
Medium	23.4 (2)	31.6 (2)	38.8 (2)	39.5 (2)	41.2 (2)
Difficult	9.4 (3)	19.0 (3)	30.7 (3)	27.0 (3)	28.7 (3)

Deletion rate 12

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy	26.7 (1)	38.7 (1)	44.4 (1)	42.1 (1)	44.5 (1)
Medium	21.0 (2)	28.5 (2)	36.6 (2)	37.8 (2)	38.5 (2)
Difficult	14.6 (3)	24.7 (3)	35.7 (3)	30.1 (3)	31.4 (3)

Table 7.2 gives the means of each test arranged by deletion rate and scoring procedure, to reveal that regardless of these variables, the three texts are always ranked in the same order by cloze. Since, therefore, deleting different words and scoring restorations in different ways has no effect on the ranking of the easy, medium and difficult texts, the null form of Hypothesis 2 is accepted.

In order to see whether the rankings reflect real differences in scores for the three texts, one-way analyses of variance were run on the texts, ignoring any differences between deletion rates. This was repeated for all five scoring procedures, and the results are shown in Table 7.3. It is evident from the results of this analysis that the differences between the texts reflected in their relative rankings are indeed real differences, and this is true regardless of the scoring procedures used.

With the obvious proviso that these results are only valid for these texts and the subjects used in this study, the clear finding is that the cloze procedure is capable of consistently distinguishing among texts for non-native speakers of English. Even if one simply counts the number of replacements which fulfil the same grammatical function as the deleted word, irrespective of whether the replacements "make sense", it is easier for non-native speakers to do this with an easy text than with a medium text, and easier with a medium text than with a difficult text. No scoring procedure, however all-embracing or permissive, seems to be insensitive to the differences in textual difficulty. Thus the cloze procedure seems to be a valid measure of text difficulty (therefore, readability) for non-native speakers of

English.

Since each subject only took one cloze test, it is impossible to compare texts to see whether they rank subjects similarly. However, it should be possible to compare the way in which different texts relate to criterion measures like the English Language Battery (ELBA) and dictation tests. Since there is no difference among groups in proficiency in English as a foreign language as determined by these measures, different correlations of cloze with the measures can be taken as providing some evidence for the possibility that different texts rank subjects differently. This will be examined in Chapter 8.

7.4.2 Scoring procedures

Hypothesis 3a applies to non-native speakers as well as to native speakers of English. One expects that different cloze scoring procedures will result in different mean scores, since each method includes more or less information than other methods.

Table 7.4 shows the results of paired t-tests on the means of each test when scored by the five different procedures. As the hypothesis predicts, the scoring procedures do, in fact, result in significantly different scores, with one exception. This exception is the comparison of the grammatically correct and the same-grammatical-function (GRCO and ACFC) procedures, for the easy and medium texts. In this case, irrespective of deletion rate, it seems that the two scoring procedures do not differ from each other in any material respect, for non-native speakers. This finding is somewhat surprising, since whilst one might expect foreigners to be able to predict the grammatical function of a cloze gap - modifier, subject, subordination marker, and

so on - it does not follow that they would be able to fill the gap grammatically; thus, although they correctly choose, say, the function of subject, they might be expected to make mistakes of concord and number. Similarly, they might correctly select a verb to fill a predicate slot, but make morphological errors in that verb (for example, choosing an incorrect past tense). Since there is indeed a difference between the same-grammatical-function (ACFC) procedure and the grammatically correct procedure (GRCO) for the difficult text, one possible explanation might be that the easier texts did not tax the reader enough - he was not induced to make grammatical errors of the type indicated by sheer incomprehension of the text. The other explanation is that the texts simply did not provide the opportunity for such errors, although this is unlikely in view of the fact that all the deletion rates (which in many cases deleted different words) produced the same result. It could be that the students were simply too good - they had no difficulties with the syntax of the two easy texts, whereas beginners or low intermediate students might have had.

The other non-significant contrasts were also with grammatical scoring procedures on the easier texts, namely, the same form class (IDFC) and the grammatically correct (GRCO) procedures at deletion rate 10 (easy text and medium text) and deletion rate 6 (medium text only).

In general, however, the expectation is confirmed that different scoring procedures contribute differently to a cloze score. The indications are, then, that different scoring procedures result in different cloze tests. However, if the rank order of the subjects remains the same, this finding is of little practical consequence.

Table 7.5 shows the Spearman rho correlation coefficients for the relationship between scoring procedures for each test, and also for each text, regardless of deletion rate differences.

The interesting result is that, unlike the results for the native speakers, there is a great deal of agreement on the ranking of subjects. Out of 120 coefficients, the lowest is .71 - i.e., 50% of the variance is accounted for by what the procedures have in common - whilst the highest is .99. In fact, 47 out of the 120 coefficients are at least .90, and many are higher.

The higher correlations seem to be achieved by the difficult text, but even the easy text, with high mean scores, gives respectably high correlations.

The closest agreement is between the same-grammatical-function and the same form class procedures (ACFC and IDFC), with rho's of the order of .98 and .99, whilst the least relationship seems to hold between the exact scoring procedure and the grammatical procedures, in particular the same-grammatical-function (ACFC), but also the identical form class procedure (IDFC). This suggests that the ability to predict grammatical function is not closely related to the ability to identify the exact word which was deleted. The former might be taken to be the lowest form of grammatical sensitivity, whilst the latter might relate more closely to, for example, sensitivity to style, awareness of author's intention, and so on. Nevertheless, even these two extreme procedures relate to one another at about the .80 level, and often higher.

The exact word method almost always (with one exception only,

on Test E12) correlates best with the any-semantically-acceptable procedure (SEMAC); the coefficients range from .85 to .95, with seven of the coefficients above .90. This close relationship is particularly marked on the difficult text, where no coefficient is below .90. It is, however, a matter of judgment as to whether the relationship is close enough for one procedure to be acceptable in place of the other, since ideally one would prefer even higher coefficients. In contrast to the native speaker study, the non-native speaker subjects are much less homogeneous in age, general ability level, interests, language proficiency and so on, and it could well be that using a more homogeneous group would result in a lower correlation between the two procedures. Nevertheless, these results confirm the findings of Stubbs and Tucker (1974) and Oller, Atai and Irvine (1974) that there is indeed a close relationship between cloze exact and cloze acceptable procedures.

Interestingly, the any-semantically-acceptable procedure (SEMAC) does not always correlate highest with the exact word method. Sometimes (Test D08) it correlates more highly with the identical form class method (IDFC), sometimes (E12) it shows a closer relationship to the same-grammatical-function procedure (ACFC), but more often it shows a closer relationship with the grammatically correct (GRCO) procedure (Tests E06, E10, M08, and M10). In other words, the semantically acceptable procedure is about as closely related to presumed measures of grammatical ability as it is to the exact word method, which may or may not be a measure of higher-order skills than those measured by the other four procedures.

Table 7.4 has shown that it is not the case that if a non-

native speaker of English correctly predicts the grammatical function of a deletion, he will tend to produce a replacement which is semantically acceptable, or, indeed, the same word as the deleted word. If this were so, the means for the different scoring procedures would show little difference.

Yet this finding of the close relationship between many of these procedures does show that a subject's performance on one measure will fairly accurately predict his performance on another measure, with the highest prediction being among grammatical scores, and the lowest between the exact word and grammatical scores.

Despite the fact that the scoring procedures agree quite highly on the ranking of subjects, it is still the case that different tests, i.e., different combinations of text and deletion rate, will result in different intercorrelations of the scoring procedures, and will also result in different rankings of the coefficients, at least for procedures which are not so closely related as the same-grammatical-function procedure (ACFC) and the identical form class (IDFC) procedure.

To summarise the findings from Table 7.5:

1) There is comparatively high agreement among scoring procedures as to ranking of non-native speaker subjects; in other words, they tend to measure the same thing, or at least in a similar manner.
2) Grammatical scoring procedures are closely related, at least one probably being superfluous.
3) The exact word method and grammatical scoring procedures show a relatively low relationship.
4) The exact word and semantically acceptable procedures are quite

closely related.

5) The semantically acceptable procedure is closely related, not only to the exact word method, but also to grammatical scoring procedures.

6) Different cloze tests display different patterns of interrelationships of scoring procedures, despite point 1 above.

Finally, as with the native speakers, a consideration of Table 7.00 shows that the rank order of scoring procedures is (easiest first): grammatically correct (GRCO), same grammatical function (ACFC), identical form class (IDFC), semantically acceptable (SEMAC), and exact word method.

In those cases (E08, E12, M06 and M08) where the grammatically correct procedure changed positions with the same-grammatical-function procedure, such that the same-grammatical-function appeared easier, the t-tests of Table 7.4 showed no significant difference between the scores, so that the rank order could have been reversed by chance alone. Only once (M12) are there significant differences among the procedures which result in a different order from that given above. On this occasion, the order is ACFC, IDFC, GRCO, SEMAC, Exact. In other words, what disagreement there is among procedures is confined to the grammatical procedures, and these procedures are in any case closely related (Table 7.5). One must conclude, as with native speakers, that whilst non-native speaker subjects may not provide replacements from the same form class as the deletion, it is more likely that they will at least provide a grammatically correct replacement. Similar to the native speakers, as the constraints on the replacement increase from mere grammaticality to semantic acceptability, and finally to include

all the constraints on the restoration of the deleted word itself, so the difficulty of correct closure increases. It remains to be seen whether this increase in constraint is linked to an increase in the amount of context available to the reader.

7.4.3 Deletion rate

Figure 7.1 shows graphically the apparent differences between means for each text at four deletion rates with each scoring procedure.

The most immediately apparent fact is that there is no consistency of direction of difference between deletion rates over different texts. In some cases deletion rate 10 is easier than deletion rate 12, in other cases it is more difficult. Similarly, sometimes deletion rate 8 is more difficult than deletion rate 10, and other times it is easier.

However, although they may not be the expected ones, some consistencies are evident. In particular, regardless of scoring procedure, the deletion rates on the difficult text reveal the same general pattern of means, namely, deletion rate 8 is always more difficult than deletion rate 6 and deletion rate 12 is always easier than 8 or 10, whilst deletion rates 6 and 12 are about the same in terms of difficulty.

The medium text reveals a somewhat opposite trend, namely, the tendency for deletion rate 12 to be more difficult than deletion rate 10 and marginally more difficult than deletion rate 6. However, in most cases, the medium text shows an increase in ease for deletion rates 8 and 10.

The easy text, however, shows a tendency for the tests to

become more difficult as the deletion rate increases, i.e., as the frequency of deletion decreases. This, of course, is the opposite of what one might expect, since common sense would suggest that as the number of words of context around a gap increases, so the difficulty of restoring the word appropriately should decrease. If nothing else, Figure 7.1 shows that the situation is by no means as simple as that.

Figure 7.1 is, however, based on raw mean scores, and does not concern itself with the question of whether the differences between the means are real differences, or whether they could have arisen by chance alone.

To answer this question, one-way analyses of variance were carried out on the four deletion rates for each text and for all five scoring procedures, and the results are presented in Table 7.6a-e.

From this table, it is clear that at least some of the apparent differences between means of the deletion rates are due to chance. In particular, the differences between deletion rates as scored by the same-grammatical-function procedure failed to reach the 5% significance level for any text.

The medium text failed to show significant differences between deletion rates with any scoring procedure, even the exact word; and the easy text failed to show significant differences among deletion rates with the semantically acceptable and grammatically correct procedures (SEMAC and GRCO), as well as with the ACFC already mentioned.

Thus, significant (at the 5% level) differences between deletion rates were found only for the identical form class procedure (IDFC) on the difficult and easy texts; the grammatically correct pro-

cedure (GRCO) and the semantically acceptable procedure (SEMAC) on the difficult text only; and the exact word method on difficult and easy texts. In other words, less than half the comparisons show significant differences.

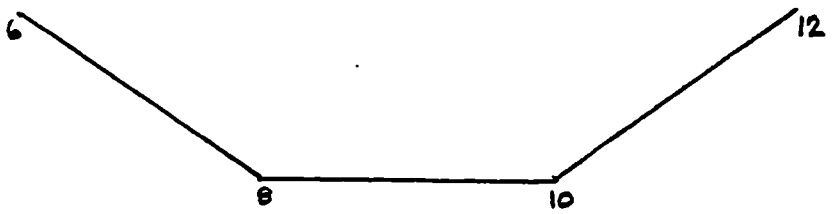
In order to see, from the six comparisons, which deletion rates are significantly different from other deletion rates, t-tests were calculated for all possible pairs on those comparisons which the analysis of variance had shown to contain significant differences. The results are tabulated in Table 7.7.

The first point to be noted is that not all pairs are significantly different - the exact word method, difficult text, has the largest number of significantly different contrasts: four out of a possible six; whilst others (exact word, easy text; semantically acceptable method, difficult text; identical form class, easy text) only show two significant contrasts.

The second point is that the significant differences are not in the same pairs of contrast. That is to say, sometimes deletion rate 6 is significantly different from deletion rate 12 (exact, easy text; IDFC, easy text), and sometimes it is not (exact, difficult text; SEMAC, difficult text). Sometimes deletion rate 8 is significantly different from deletion rate 6 (GRCO, difficult text; exact, difficult text), sometimes it is not (exact, easy text; SEMAC, difficult text). The sole consistent result is that deletion rate 8 is never found to be significantly different from deletion rate 10.

The third, and most important, result to be noted is the direction of the significant differences that <u>were</u> discovered.

For the exact word method, difficult text, deletion rates 8 and 10 were not different, nor were deletion rates 6 and 12, but both deletion rate 6 and deletion rate 12 produced higher mean scores than deletion rates 8 and 10. If similar positions on the horizontal are taken to mean "no significant difference", and a difference in positions is taken to show that significant differences existed, then the situation can be represented graphically as:



For the easy text, however, the position is quite different, i.e.,

which can be seen to represent a situation where deletion rates 6, 8 and 10 are not different, whereas deletion rate 12 is lower, that is, more difficult, than deletion rates 6 or 10. Unfortunately, this figure misrepresents the situation to some extent, since deletion rate 8 is not significantly different from deletion rate 12.

Equally difficult to represent graphically is the situation with the semantically acceptable scoring procedure (SEMAC), difficult

text, where deletion rate 10 is significantly more difficult than deletion rates 12 or 6, but not significantly different from deletion rate 8. These two differences are the only significant differences for this scoring method.

The identical form class method (IDFC), difficult text, reveals a situation similar to that of the exact word method, viz., that deletion rate 6 is significantly different from and easier than deletion rates 8 and 10, but not deletion rate 12, and deletion rate 8 is significantly more difficult than deletion rate 12. Here, however, not only is there no difference between deletion rates 8 and 10, and deletion rates 6 and 12, but the comparison 10:12 shows no difference either.

On the easy text, scored by the identical form class method, the only differences to be uncovered were deletion rate 12 with deletion rate 6, and deletion rate 12 with deletion rate 10. In both cases, deletion rate 12 proved the more difficult of the pair.

Finally, for the grammatically correct procedure, difficult text, deletion rate 12 was easier than deletion rate 10 only, whereas deletion rate 6 was easier than deletion rates 8 and 10, but not different from deletion rate 12.

Thus, it is at least clear that there is no clear trend in terms of consistent differences between deletion rates, regardless of text differences. However, there are greater consistencies if one takes one text at a time.

Nevertheless, even the apparent tendency (from an inspection of the means alone) for the difficult text, regardless of scoring

procedure, to result in lower mean scores at deletion rates 8 and 10 than at deletion rates 6 and 12, is only consistent in that deletion rate 6 is always easier than deletion rate 10. It is not, however, always easier than deletion rate 8. Although deletion rate 12 is never easier or more difficult than deletion rate 6 (for the difficult text), it is only sometimes easier than deletion rates 8 or 10.

The situation is more regular with regard to the easy text. Irrespective of scoring procedure (IDFC or Exact), deletion rates 6, 8 and 10 are never different, whereas in both cases both deletion rate 6 and deletion rate 10 are <u>easier</u> than deletion rate 12.

Overall generalisations are thus clearly impossible, since with one text certain deletion rates produce easier tests, whereas with another text, contradictory results are achieved. It is hard to conclude, therefore, that the differences between tests are due to the differing number of words surrounding each item, for, if this were so, one would expect both consistency (probably irrespective of text) and greater context to result in easier tests. Instead, as the difficult text shows, the two most different deletion rates (6 and 12) result in apparently similar tests in terms of difficulty.

We must look elsewhere to explain the difference between cloze test forms, and the answer is probably to be found in the simple fact that different words are deleted on each deletion rate. However, in order to confirm that this is so, it is necessary to compare the cloze deletion rates, looking at only those items which are common to both deletion rates under consideration.

(As was explained in Chapter 5, the counting for deletions

in any text always began at the same word; consequently, the second word deleted at deletion rate 6 is the same word as the first word deleted at deletion rate 12, and the fortieth word deleted at deletion rate 10 is the same word as the fiftieth word deleted at deletion rate 8. It is therefore possible to compare deletion rates based on only those items common to the deletion rates under consideration.)

The same computer programs used in the native speaker study were used to select the common items, calculate means and standard deviations and perform t-tests for the differences between the means. These calculations were done only for those scoring procedures and texts for which the analysis of variance (Table 7.6) had shown significant differences between deletion rates. The analysis was therefore carried out on the difficult text, scoring procedures Exact, SEMAC, GRCO and IDFC; and the easy text scored by the exact and IDFC methods. The results are presented in Table 7.8.

The clear result of this analysis is that no significant differences between deletion rates were found. The sole exception to this is the comparison between deletion rates 8 and 10, exact word method, difficult text, and this is unimportant for two reasons. Firstly, as Table 7.7 shows, the contrast of deletion rates 8 and 10 for this score on this text is not significant when all the items are considered, not just the identical items. Secondly, the means and distributions are so low that it is doubtful whether the t-test is even applicable to these two sets of scores.

For all other comparisons, whatever the text, whatever the scoring procedure, no significant effect was found of increasing the

amount of context around a cloze deletion.

To summarise the results from Table 7.6, 7.7 and 7.8:

1) Despite apparent differences in mean scores, no significant differences between deletion rates were found for 60% of the texts and scoring procedures studied.
2) In particular, the scoring procedure ACFC (same grammatical function) never showed significant differences between deletion rates.
3) No significant differences between deletion rates were found on the medium text, regardless of scoring procedure.
4) From all the possible contrasts of deletion rate pairs for those texts and scoring procedures that showed significant differences somewhere, 44% were found to be significantly different.
5) Deletion rates are not consistently different from each other - for any possible pair of deletion rates, sometimes they are different, sometimes they are not.
6) Where there is a difference among deletion rates, its direction varies. There is a tendency for deletion rates 8 and 10 to be more difficult than the other two on the difficult text, and for deletion rate 12 to be more difficult than the rest on the easy text. However, these differences are a) not as predicted, and b) not consistent across texts.

Points 1 to 6 above can be further summarised:

Where there are differences among deletion rates on cloze tests, they are not consistent.

7) When only identical items were considered, no significant effect of varying the deletion rate was found, which can be interpreted to

mean that varying the amount of context around a cloze item has no effect on its difficulty.

7.4.4 The efficiency of the cloze test

This section will deal with the efficiency of the cloze test with non-native speaker subjects, and the influence on this efficiency of the three main variables: text, deletion rate and scoring procedure. This efficiency will be examined internally in this chapter, whilst the next chapter (Chapter 8) will be concerned with, amongst other things, cloze as a test of English proficiency as compared with external criteria, from which a limited comparison of efficiency will be possible.

In relation to the internal criteria, particular reference will be made to the descriptive statistics of the tests, the reliability measures, and a traditional item analysis. The remarks in Chapter 6 apply here insofar as they refer to the difficulty of defining the desired efficiency for a test whose purpose one does not clearly know. However, one can perhaps more readily conclude here that for non-native speakers the cloze test can be considered to be used as a norm-referenced proficiency test, since it is precisely with such tests that it is being compared. Thus it can plausibly be expected that a wide distribution would be preferred to a narrower one, and that, other things being equal, a mean of around 50% would be preferred to more extreme means. Yet, as already indicated in Chapter 6, Bormuth (1968a) seems to suggest that a mean of around 44% might be equivalent to a multiple-choice comprehension score of 75%, which is felt to indicate the study level of comprehension with native speakers. Therefore one might well expect the cloze with non-native speakers consistently to

return lower mean scores than more traditional tests. The native speaker study (Chapter 6) has already confirmed this to the extent that it showed no subject achieving over 90% exact restorations, and that even the easiest text with the exact word scoring procedure still resulted in means of only 70%.

With these non-native speakers the highest mean scores were achieved with the most liberal scoring procedures on the easiest texts, and these were of the order of 90% (Table 7.00). These grammatical procedures gave lower means on the medium text (from 80% down to 73%), but in fact even on the difficult text mean scores were still above 50% (ranging from 51% to 75%). Not surprisingly, as the means increase from difficult to easy text, with the grammatical scoring procedures, the dispersion decreases, so that whereas on the difficult text good distribution is achieved (standard deviations between 20% and 35% of the mean), on the easy text this is reduced to around 10% of the mean (Table 7.00B). Of course, this is partly a function of the increase in mean scores, but also reflects an absolute decrease in standard deviation, down to about 8 percentage points from as high as 20. Nevertheless, the GRCO procedure results in the absolutely - as opposed to relatively - highest standard deviation of all, and in general the distribution achieved by grammatical scores is reasonably large. Interestingly, only on the easy text do even the grammatical scoring procedures result in maximum (100%) scores, so the ceiling effect is not a large problem.

The any-acceptable (SEMAC) procedure has a different effect on each text. On the easy text, although SEMAC mean scores are lower than grammatical means, they are not greatly so - between 6 and 11 per-

centage points only - and the mean remains at about 80%. The difference on the easy text lies in the distribution of scores, since the standard deviations of the SEMAC are considerably higher than those of the grammatical scoring procedures, and even those of the exact word scoring procedure.

On the medium text, the SEMAC results in means of about 60%, which are considerably lower than the grammatical score means - by 14 to 18 percentage points. Here again the largest absolute distribution of all scoring procedures is achieved by the SEMAC.

The same remark applies almost always to the difficult text, where the SEMAC distribution is again large. On this text even greater differences in mean scores are evident, so that the SEMAC, achieving around 40-50% means, is midway between the exact and grammatical scoring procedures (about twenty percentage points different from both). The distributions achieved by this procedure, together with the moderately high means, indicate that it is efficient at discriminating subjects, and it is at an appropriate level of difficulty on the medium and difficult texts.

The exact word procedure results in much narrower distributions, on both medium and difficult texts, than any of the other procedures. On the difficult text this could be attributed to the low means - 19-29%. However, on the medium text, where reasonable means are evident (40-45%), the discrimination remains poor. Only on the easy text, with means of around 60%, is the distribution comparable with that of other scoring procedures. In fact, the standard deviation for the exact procedure remains remarkably constant, regardless of text change.

As in Chapter 6, the results with non-native speakers have shown that the difference between deletion rates appears to be due to the mere fact that different words have been deleted, and not to the increase or decrease in availability of context. Although there are great differences in the descriptive statistics of different deletion rates, there seems to be little point in comparing deletion rates for efficiency. The difference between texts is clearly constant, and so a comment on the interaction between text and scoring procedure as it affects the efficiency of the cloze seems to be in order.

An interesting point is that although the exact word method achieves reasonable mean scores on the medium and easy texts, the distribution does not change shape, and is in no case the best. Since better distribution is almost always achieved by the SEMAC procedure, and the mean scores are still within acceptable limits, it seems to be preferable to the exact score on both medium and difficult texts. As neither maximum (100%) nor minimum (0%) possible scores are attained by any subject on this procedure, the actual minimum and maximum range is from 4 to 42 (8% to 84%), an encouragingly large spread. Thus, either the medium or the difficult text, providing they are scored by the SEMAC rather than the exact procedure, would appear to be suitably efficient, at least in as far as an inspection of the descriptive statistics reveals test efficiency.

In terms of reliability, three measures were taken - the standard error of the mean, the standard error of measurement, and the KR20 reliability coefficient. The latter, although a measure of internal consistency, is something of an underestimate if the spread of

item difficulties is uneven.

The standard error of the mean is notably lower for the exact score on the difficult text than for other scores (Table 7.00), but this is because the exact mean is low in any case, and, expressed as a percentage of the mean (Table 7.9a), it is, in fact, always higher for the exact score than for any other. Expressed in relative terms, the standard error of the mean is lowest for the grammatical scoring procedures. Table 7.9b sets out a comparison of the standard error of measurement for the exact and SEMAC scoring procedures, from which it is clear that the standard error of measurement is always proportionately lower for the SEMAC than the exact method, although in absolute terms it is only lower than that of the exact scoring procedure on the easy text. The difference between the two procedures in absolute terms is, however, minimal, giving the edge to the SEMAC because of its better performance on the relative reliability. The lowest figures, absolute and relative, are gained with the easy text.

Some variation is evident in the KR20 reliability coefficients, although a figure as low as .53 (exact, E6) is exceptional. The grammatical and SEMAC procedures range from .74 to .92, with the SEMAC tending to achieve marginally better reliability than the grammatical procedures. In virtually every case the exact procedure results in lower reliability than the other procedures, with a greater range of coefficients. No consistent advantage is evident for one text over another, even when only one scoring procedure is considered.

The conclusion of this study of the reliability measures, then, is negative, in that no one text consistently gives more reliable

results than another. The study also shows the exact word procedure to perform less reliably than the others. With the remaining scoring procedures, however, the reliability, although variable, appears acceptable.

Incidentally, it is at least noteworthy that different deletion rates on the same text with the same scoring procedure give reliability estimates which vary at least as much as do the differences between texts or scoring procedures (KR20 exact D10 = .69, D12 = .80; exact E6 = .53, E12 = .80). As expected, it is not the case that as amount of context increases, the results become more reliable; it is, however, true that as the deletion rate changes, so, unpredictably, will the reliability. This, moreover, applies regardless of scoring procedure.

Table 7.10 presents a summary of the item difficulties for all 12 tests scored by the five different procedures.

Items with greater than 80% or less than 20% facility are traditionally considered to be so extreme as to detract from the efficiency of a test. Using this criterion, the grammatical scoring procedures are clearly much less efficient than the exact or SEMAC on all three texts, but particularly on the easy and medium texts. No consistent differences are evident among deletion rates for these three grammatical procedures. Although they rarely result in items that no subject gets correct, they do provide items which everyone answers correctly, especially on the easy text. Nevertheless, on the difficult text a fairly high proportion of items, even with the grammatical scoring procedures, comes within the above-mentioned limits of acceptability (from 58 to 78% of items, in fact). On this text, this is considerably better

than the exact word procedure, which only has between 30% and 48% of items within the same limits, and between 44% and 66% with a facility index of less than 20%. Of course, the exact word procedure improves in efficiency on the medium text (between 50% and 66% within acceptable limits) and the easy text. Nevertheless, even on the easy text the maximum number of acceptable items is 66% of the total (E6), and in one case (E10) the number is below 50%. Similarly, even on this text, one item per test proved so difficult that nobody was able to restore it correctly. It is, however, true that on the easy text the exact procedure results in the best distribution of item difficulty of the five procedures.

In contrast, the SEMAC performs relatively poorly on the easy text, resulting in a preponderance of easy items. On the medium text the SEMAC results in fewer extreme items than any other procedure, whilst on the difficult text it is also the best procedure in terms of numbers of efficient items. The interesting thing is that on the difficult text the SEMAC is not much more efficient than any grammatical procedure, with the exception of Test D6.

From this detailed survey, it is at least clear that all the cloze tests and all scoring procedures frequently result in inefficient items, since the lowest percentage of inefficient items (18%) is achieved once only, whereas a proportion of over 80% of items proving to be inefficient is attained ten times.

By comparing the ratio of efficient to non-efficient items across tests and scoring procedures, it is possible to get an idea of the most and least efficient tests (Tables 7.11, 7.12, and 7.13).

T A B L E 7 . 1 1

Performance of scoring procedures, in terms of item difficulty. Non-native speakers.

	Best	Worst
D6	SEMAC	Exact
D8	SEMAC/ACFC	Exact
D10	GRCO	Exact
D12	SEMAC	Exact
M6	SEMAC	ACFC
M8	SEMAC	ACFC
M10	SEMAC	ACFC
M12	SEMAC	ACFC
E6	Exact	GRCO
E8	Exact	GRCO
E10	Exact	ACFC
E12	Exact	ACFC

In terms of scoring procedures, no one procedure emerges as consistently more efficient than the rest, since the SEMAC is best on the difficult and medium texts, whilst the exact procedure is best on the easy text. On the other hand, the exact method is worst on the difficult text, whilst on the other two texts grammatical procedures are the least efficient (Table 7.11).

The deletion rates perform inconsistently across scoring procedures and texts, so that no one deletion rate can be recommended for

efficiency (Table 7.12).

T A B L E 7 . 1 2

Best performance of deletion rate, in terms of item difficulty. Non-native speakers.

	Exact	SEMAC	GRCO	IDFC	ACFC
Easy text	8	12	8	12	8
Medium text	12	6	12	6	6
Difficult text	6	6	8	10	8

T A B L E 7 . 1 3

Best performance of text, in terms of item difficulty. Non-native speakers.

Deletion rate	Exact	SEMAC	GRCO	IDFC	ACFC
6	E	D	D	D	D
8	E	D	D	D	D
10	M	D	D	D	D
12	M	D	D	D	D

On the other hand, very consistent results are seen in Table 7.13, where the difficult text emerges as the most efficient text for use with all deletion rates and scoring procedures except for the exact scoring procedure, where both easy and medium texts are more efficient

than the difficult text.

In summary, then, in terms of non-native speakers, if the exact word procedure is being used, then it should probably be with an easy text, but the deletion rate used would appear to be unimportant. If a difficult text is being used, then it should definitely not be scored by the exact method, but rather by a grammatical procedure, or, preferably, SEMAC. If the SEMAC is being used, then either a medium or difficult text would produce a relatively efficient test.

Finally, Table 7.14 compares item discrimination across tests for both the exact word and the SEMAC scoring procedures.

From this it can be seen that the exact word procedure results in greatest negative discrimination on an easy text, and the highest number of acceptable items tends to be achieved with the medium text. It should be noted, however, that only once does this procedure produce more than half of its items with acceptable (above + 0.2) discrimination (Test E8). The number of items with both acceptable facility and acceptable discrimination ranges from a mere 20% to 46%. The SEMAC results in considerably better item discrimination indices. Not only is the number of negatively discriminating items greatly reduced, to virtual insignificance, there is also a higher proportion of items with reasonable positive discrimination, as compared with the exact method (from 40–82%). This scoring procedure also results in a higher proportion of items with both acceptable facility and discrimination (from 26–72%). On both medium and difficult texts the SEMAC invariably results in better item discrimination, and even on the easy text it produces fewer negative discriminations and approximately equal all-

round acceptability.

In terms of discrimination, then, the cloze seems to be relatively poor at discriminating non-native speakers, but the SEMAC produces better discrimination than does the exact word scoring procedure.

CHAPTER 8

Results 3:
Cloze as a Measure of Proficiency in English as a foreign language

The null form of Hypothesis 4 states:

a) There is no difference between deletion rates as measures of English-as-a-foreign-language proficiency.

b) There is no difference between texts as measures of English-as-a-foreign-language proficiency.

c) There is no difference between scoring methods as measures of English-as-a-foreign-language proficiency.

d) There is no interaction between deletion rates, text and scoring method as predictors of English-as-a-foreign-language proficiency.

This hypothesis will be investigated by looking at the relationship between the cloze tests and the ELBA results. The dictation tests, which were also administered to most of the subjects, are in a sense experimental, and so not to be considered established measures of proficiency in English as a foreign language, whatever recent theoretical claims may have been made in their favour. The relationship between cloze and dictation, and indeed between dictation and ELBA, will be examined later in this chapter.

Before proceeding to an examination of the above hypothesis, it is proposed to examine the results of the ELBA as administered to the subjects of this study, and to compare them with the results reported in the test manual, and in Chapter 5 of this study.

8.1 Results of the ELBA Test

The intercorrelations (internal validity coefficients) of the ELBA tests and subtotals with one another and the total score are given in Table 8.1

TABLE 8.1

Intercorrelations of the subtests of the English Language Battery (ELBA) (Pearson Product Moment correlations)

	1	2	3	4	5	6	7	Pt 1	Pt 2	Total
Test 1		.37	.33	.62	.61	.60	.47	.91	.64	.80
2	.37		.45	.51	.56	.42	.47	.60	.54	.59
3	.33	.45		.38	.45	.23	.37	.53	.39	.47
4	.62	.51	.38		.77	.76	.71	.85	.84	.89
5	.60	.56	.45	.77		.75	.69	.77	.93	.91
6	.60	.42	.23	.76	.75		.62	.72	.92	.87
7	.47	.47	.37	.71	.69	.62		.65	.79	.77

n = 264

(all coefficients significant at the 0.1% level)

These are, on the whole, lower than those reported by Ingram in the ELBA Manual, but are, of course, based on a different and slightly smaller group of subjects. According to one interpretation, they are more satisfactory than the original coefficients, since the test intercorrelations are lower and thus indicate less redundancy in the information provided by each test. However, they confirm the

evidence from the manual that Tests 4, 5, 6 and 7 (Listening Comprehension, Grammar, Vocabulary, and Reading Comprehension respectively) are all quite highly related. This is almost certainly due to the heavy reading requirement for all four tests, including the Listening Comprehension Test. In fact, the Listening Comprehension Test is much less closely related to the other tests of listening than to the reading tests. Tests 2 and 3 (Intonation and Stress) are relatively unrelated to the overall total (.59 and .47), whereas the other five tests appear to be closely related to the Total. However, when the individual tests are correlated with the Total minus themselves (Table 8.2), Test 1 (Sound Recognition) also reduces in importance as a predictor of such a total. Even Reading Comprehension (Test 7) is less important that Tests 4, 5 and 6. The reading comprehension required for Test 7 is probably not the same as that required for or explicitly tested by Tests 4, 5 and 6. In particular, some 75% of the questions are inferential, rather than referential, whilst well over half of them require reference to intersentential connections and relations rather than sentential relations.

After establishment of the correlation matrix, the individual ELBA tests were factor-analysed according to the principal components factor extraction method, with unities in the main diagonal. The factors thus extracted were subjected to two rotations using the Varimax procedure (which assumes orthogonality). The first solution attempted to rotate all factors with an eigenvalue of at least 1.0. However, Ohnmacht, Weaver and Kohler (1970) express dissatisfaction with such an arbitrary cessation of factoring, and they themselves present several

solutions for consideration, including factors with eigenvalues lower than 1.0. Clearly, the solution one adopts depends on one's purpose and desire for parsimony, but they present evidence that a more complete and satisfying solution is obtainable by going below EV 1.0, especially if the number of tests in the matrix is small. For this reason, a second rotation was undertaken, including all factors with an eigenvalue of at least 0.5. This, of course, has the effect of increasing the communalities, and gives a more complete solution. Which solution one chooses to adopt is purely a matter of judgment, but, in the case of the ELBA data, more information certainly results from the second solution. The two solutions are presented in Table 8.3 - for all subjects, and also for three groups of approximately 90 subjects each, divided according to the cloze text taken. This procedure is necessary, since the factor analysis of cloze, dictation and *ELBA must be based on each cloze text, instead of* all cloze texts grouped together. Thus, it is helpful to be able to compare the analysis of cloze, dictation and ELBA with an ELBA-only analysis for each cloze text group.

TABLE 8.3

Factor analysis of ELBA, overall and for three subgroups.

Rotation = Varimax

Overall

Eigenvalue $\geq$ 1.0

Test (communalities of 1, 2, 3 $<$ 0.6)

1	.75479
2	.69052
3	.63998
4	.89753
5	.90794
6	.82365
7	.80587

Eigenvalue $\geq$ 0.5

Test (no communality $<$ .75)

1	.01361	.85230	.32174
2	.74775	.11873	.48113
3	.23835	.22969	.89461
4	.52818	.71743	.19216
5	.48590	.72853	.26159
6	.40154	.80362	.01704
7	.70397	.49939	.11038

Difficult Text Subgroup

Eigenvalue $\geq$ 1.0

Test (communalities of 1, 2, 3, 7 $<$ 0.6)

1	.69927
2	.72287
3	.54947
4	.85529
5	.89595
6	.79510
7	.71839

Eigenvalue $\geq$ 0.5

Test (communalities $>$ .73)

1	.31281	.86238	.06780	.21305
2	.16455	.60382	.68516	.01986
3	.18351	.16255	.11136	.95734
4	.71858	.36160	.33603	.12653
5	.75476	.25350	.36550	.28795
6	.91489	.14826	.15401	.09494
7	.37484	.01222	.84188	.15206

Medium Text Subgroup

Eigenvalue $\geq$ 1.0

Test	(no communality $<$ .6)	
1	.80052	-0.00662
2	.12353	.85265
3	.18838	.79919
4	.78886	.41605
5	.81218	.34171
6	.82920	.15219
7	.63336	.53254

Eigenvalue $\geq$ 0.5

Test	(no communality $<$ .68)		
1	.48166	.13312	.77624
2	.41108	.68999	-0.41977
3	.10732	.91293	.25054
4	.81200	.32808	.19355
5	.81762	.25738	.22586
6	.88037	.01146	.13524
7	.68127	.45767	.11728

Easy Text Subgroup

Eigenvalue $\geq$ 1.0

Test	(2 and 3 $<$ .6 communality)
1	.84052
2	.75069
3	.71102
4	.93294
5	.94618
6	.87015
7	.84635

Eigenvalue $\geq$ 0.5

Test	(no communality $<$ .73)	
1	.85117	.23705
2	.33748	.84991
3	.26262	.89110
4	.84969	.40665
5	.82752	.46408
6	.91020	.20157
7	.77694	.35968

Taking the overall result first, the first solution (EV $\geq$ 1.0), only one factor emerges. The variable loadings are high for Tests 4, 5, 6 and 7, and somewhat lower for 1, 2 and 3. This factor must be interpreted as a general English-as-a-foreign-language

proficiency factor, with an emphasis on reading. According to this view, English-as-a-foreign-language proficiency as measured by ELBA is one unitary, generalized ability.

However, on the second solution (EV $\geqslant$ 0.5), three factors emerge, and, of course, the communalities are higher. The first factor loads on 2 and 7, but only moderately on 4, 5 and 6. The second factor loads on 1, 4, 5 and 6, and moderately on 7, whilst the third factor loads on 3, with minor loadings on 1 and 2. The identification of these factors is somewhat problematic. The third factor might be called "listening", specifically suprasegmental. The second factor appears to be a more general English-as-a-foreign-language proficiency factor, irrespective of mode of presentation of material. The first factor, however, is interesting in that it loads on a listening variable and a reading variable. However, in reality, when one looks closely at Test 2, it is clear that there is a high reading requirement, especially since the test is timed. Subjects have to read quickly and relate what they have read to what they hear. They also have to infer emotions, moods and attitudes from the intonation and stress pattern, and relate these to the written gloss.

The factor analysis for the cloze text groups gives a broadly similar first solution (EV $\geqslant$ 1.0) on the difficult and easy texts - i.e., the solution is unifactorial, with somewhat lower loadings on variables 2 and 3. The medium text gives two factors, the first one loading on Tests 1, 4, 5 and 6, and loading lower on 7, with the second factor loading on 2 and 3, with some loading on 7. This second factor is reminiscent of Factor 1 on the overall ELBA (EV $\geqslant$ 0.5).

The second solution (EV $\geqslant$ 0.5) is different for each group. The difficult text group is the most complex, with four factors. Factor 1 loads on 4, 5 and 6; Factor 2 on 1, with some 2; Factor 3 on 7 and some 2; Factor 4 on 3 alone. This Factor 3 is again reminiscent of Factor 1 on the overall ELBA (EV $\geqslant$ 0.5) and Factor 2, medium text (EV $\geqslant$ 1.0). The first factor is broadly similar to Factor 2, overall ELBA (EV $\geqslant$ 0.5).

On the medium text, three factors emerge. The first again loads on 4, 5 and 6, with some 7. The second factor loads on 3 (with some 2) and the third loads on Test 1 almost exclusively. The easy text, on the other hand, reveals only two factors with an eigenvalue greater than 0.5. The first loads on 1, 4, 5 and 6, with some 7, and the second loads on 2 and 3, with much lower loading on 4, 5 and 7.

The upshot of this comparison of overall and different subgroups is twofold. Firstly, taking a traditional conservative approach, ELBA appears to be unifactorial, even for the subgroups. Tests 2 and 3, loading lower than the rest on this factor, are somewhat apart from the other tests. When a more complete solution is sought, the monolithic English-as-a-foreign-language proficiency factor breaks up into a number of other factors. These factors are not consistent across subgroups of the population, and labelling them is also somewhat problematic. A clear grouping is achieved of Tests 4, 5 and 6 (Factor 1, all subgroups; Factor 2, overall). Test 1 sometimes loads on the same factor (Factor 2, overall; Factor 1, easy text) and sometimes loads on its own factor (Factor 2, difficult text; Factor 3, medium text). Test 3 loads on its own factor (Factor 3, overall and difficult text) or in conjunction

with Test 2 (Factor 2, easy text and medium text). The reading comprehension test (7) is odd in that it tends not to load high on its own factor, but rather to be associated with other factors, loading somewhat lower than 4, 5 and 6 on the same factors (Factor 2, overall; Factor 3, difficult text).

One clear conclusion from this is that a division of ELBA into two parts labelled "Reading" and "Listening" is not justified by this analysis. The difference between tests is less one of mode (written versus spoken) than of content (linguistic/metalinguistic), or scope (sentential/suprasentential). In this way, reading comprehension (Test 7) is different from listening comprehension (Test 4) not because of the mode, but perhaps because of the nature of the skills required: the ability to make inferences, perhaps, or to handle text rather than sentences. The most consistently present factor loads on 4, 5 and 6 mainly, and might be termed core proficiency, since it covers ability to deal with the syntax and lexis of a language, and to comprehend (and read) its sentences.

8.2 Cloze and ELBA

As the results of the previous chapters have shown, it is misleading to ignore the difference between cloze texts in order to examine deletion rates; thus, section a) of Hypothesis 4 must be examined for each text. However, to establish possible differences between texts in their relationship with a recognised measure of English-as-a-foreign-language proficiency, it would appear valid to ignore the differences between deletion rates, and aggregate the results to give three main cloze text/tests - easy, medium and difficult - which one

can correlate with ELBA. To investigate the nature of different scoring procedures as measures of English-as-a-foreign-language proficiency, it is possible to examine the correlation coefficients both for each test and for each text regardless of deletion rate.

8.2.1 Deletion rates

Table 8.4 gives the Pearson Product Moment Correlation Coefficients of each of the twelve cloze tests, scored by five different scoring procedures, with the ELBA tests, subtotals and total scores. For the purpose of investigating the hypothesis, the ELBA Total will be taken to be an aggregate and adequate measure of overall English-as-a-foreign-language proficiency.

If we take the tests text by text to compare deletion rates, it is clear that different deletion rates give different correlation coefficients. The exact word scoring method, difficult text, correlates quite high with the ELBA Total (D08, .82; D10, .79; D12, .77), but not always - deletion rate 6 shows a much lower correlation at .51. Similarly, on the same text, the any-acceptable-word method shows generally high coefficients (D08, .87; D10, .83; D12, .85) but with an exception at deletion rate 6 (.67).

The exact word method, medium text, shows greater disagreement among deletion rates. This time, the lowest coefficient is with deletion rate 10, at .57; the next lowest with deletion rate 8 at .68; and the highest with deletion rate 6, at .86. The any-acceptable-word procedure (SEMAC) shows a similar pattern, but reveals greater agreement among deletion rates: M10, .74; M08, .77; M12, .78; M06, .88.

However, on the easy text the best correlations with ELBA

were deletion rate 8, for the exact word method (.70), and deletion rate 12 for the SEMAC procedure (.77). In this case, the any-acceptable procedure shows little variation amongst the coefficients (E06, .74; E08, .69; E10, .74; E12, .77), whereas the exact word method shows greater differences (E06, .59; E08, .70; E10, .65; E12, .67).

In general it seems that different deletion rates do indeed result in different correlations with a measure of proficiency in English as a foreign language, regardless of text, at least for the exact word method. The nature of the difference is, however, unpredictable, since with one text deletion rate 8 shows the lowest correlation, with another text it has the highest correlation with ELBA. Similarly, deletion rate 6 had a relatively low correlation with ELBA on the difficult text, but on the medium text it showed a much closer correlation. It is thus possible to reject the null form of Hypothesis 4a), and conclude that different deletion rates may very well result in different correlations with a measure of overall English-as-a-foreign-language proficiency. It is, however, also possible to relate these findings to Hypothesis 4d), and show that there is an interaction between deletion rate and scoring method at least, since it is clear that using the any-acceptable-word procedure results in greatly reduced differences between correlation coefficients.

In other words, one way to stabilise cloze/ELBA coefficients would be to use the any-acceptable-word scoring procedure instead of the exact word method.

It is possible, when examining the effect of different deletion rates on the same text with the same scoring procedure, to look

at the correlation of the cloze with the ELBA subtests, to see whether a change of deletion rate results in different skills being measured by the cloze, or rather, and more likely, whether the skills being measured relate differently to different deletion rates.

Just as the magnitude of the correlation with the ELBA Total varies, so, as Table 8.4 shows, does the size of the correlations of cloze with individual tests. Indeed, sometimes the test correlates significantly with cloze, and sometimes the correlation is not significantly different from zero. Thus, it is no more possible to predict the size of the correlation of cloze with an individual test than it was to predict the size of the correlation of cloze with the overall ELBA Total.

It is conceivable, however, that within this instability, there is a stability of order of importance of the individual tests which would suggest that the same underlying skills are being measured by each cloze test, regardless of the magnitude of the correlation.

Table 8.5 gives, for each cloze test, scored by the exact word method and the any-acceptable word method, the ELBA tests in order of size of correlation with cloze.

TABLE 8.5

Rank orders of correlations of ELBA tests with cloze, by cloze test. Exact word and any-acceptable-word scoring procedures.

	ELBA Test								ELBA Test						
Cloze	1	2	3	4	5	6	7	Cloze	1	2	3	4	5	6	7
Exact								SEMAC							
D06	NS	2	NS	1	3	NS	NS	D06	3	1	NS	2	NS	NS	NS
D08	6	5	NS	3	1	2	4	D08	6	5	NS	3	1=	1=	4
D10	NS	5	NS	3=	2	1	3=	D10	6	5	NS	3	1=	1=	4
D12	3	7	6	4=	1	2	4=	D12	3	NS	6	4	1	2	5
M06	5	6	7	1	2	4	3	M06	4=	6	7	1	3	2	4=
M08	1	NS	5	3	2	4	6	M08	6	7	2	1	3	5	4
M10	3	2	NS	NS	1	NS	4	M10	1=	3=	6	5	3=	NS	1=
M12	NS	4	NS	2	3	1	5	M12	5	4	7	3	2	1	6
E06	5	NS	NS	3	2	4	1	E06	5	7	6	3	1	4	2
E08	4	5	NS	1	2	3	6	E08	6	4	5	1	2	3	7
E10	3	NS	NS	1	2	4	5	E10	3	7	6	2	1	5	4
E12	6	1	NS	3=	3=	2	5	E12	6	1	7	3	2	5	4

NS Not significant

From this table, no such stability emerges. When cloze is scored by the exact word method, different cloze tests associate with different ELBA tests. In fact, all of the ELBA tests except one are more closely associated with at least one cloze test than the other ELBA tests are. Thus, Sound Recognition shows the highest correlation for cloze Test M08; Intonation correlates more highly than any other ELBA test, with cloze Test E12; Reading Comprehension is the highest for Test E06, Vocabulary for Tests D10 and M12, Grammar for Tests D08, D12 and M10, and Listening Comprehension for Tests D06, M06, E08 and E10. The only test that does not, at some point or other, correlate highest with cloze, is Test 3, Sentential Stress.

Nevertheless, a tendency is discernible for Tests 4, 5 and 6 (Listening Comprehension, Grammar, and Vocabulary) to be more closely associated with cloze than the other tests, and for Tests 2 and 3 (Intonation and Sentence Stress) to be least associated with cloze.

Similar conclusions must be drawn about the different relationships of the individual ELBA tests with cloze as scored by the any-acceptable-word procedure, with the (probably unimportant) difference that Sound Recognition never correlates highest with cloze.

Because of the low numbers of subjects taking any one cloze test and ELBA, the reliability of these results is, of course, somewhat suspect, and it is possible that at least some of the fluctuations in correlations shown by the data may be due to the small numbers of people involved rather than to characteristics of the tests themselves. Even bearing this in mind, it seems reasonable to conclude, however, that there is little stability in the association of individual cloze tests

with any of the individual ELBA tests, whatever the scoring method. Thus, one must be extremely careful when making statements about "what cloze tests" to make clear that this would not necessarily hold if a different deletion frequency were used to produce the cloze test.

8.2.2 Text

To examine the difference between texts in their correlation with ELBA, it is possible to compare coefficients for each text, holding deletion rate constant. The relevant data from Table 8.4 is recast, for ease of inspection, in Table 8.6, for each scoring procedure.

T A B L E 8 . 6

Correlation of ELBA Total with various cloze tests

	TEXT		
	Easy	Medium	Difficult
Deletion Rate 6			
Exact	.59	.86	.51
SEMAC	.74	.88	.67
GRCO	.60	.81	NS
IDFC	.44	.67	.43
ACFC	.45	.68	.43
Deletion Rate 8			
Exact	.70	.68	.82
SEMAC	.69	.77	.87
GRCO	.61	.74	.73
IDFC	.50	.51	.80
ACFC	.46	.50	.74
Deletion Rate 10			
Exact	.65	.57	.79
SEMAC	.74	.74	.83
GRCO	.75	.75	.79
IDFC	.63	.70	.83
ACFC	.65	.65	.82
Deletion Rate 12			
Exact	.67	.73	.77
SEMAC	.77	.78	.85
GRCO	.72	.75	.68
IDFC	.73	.70	.72
ACFC	.71	.69	.70

From this table, it is clear that the difficult text has the closest relationship with the ELBA Total, on three of the four deletion rates. However, for deletion rate 6, it appears that the medium text results in higher correlations than the difficult text, regardless of scoring procedure. Interestingly, the clear superiority of the medium text at this deletion rate is greatly reduced when the any-acceptable-word scoring procedure is used instead of the exact word method. In fact, the any-acceptable-word method generally tends to minimise the differences between texts, in terms of correlation with ELBA, whereas the exact word procedure tends to emphasise the differences. Thus, not only does it appear that the null form of Hypothesis 4b must be rejected, since clearly different correlations are achieved by different text, but also further evidence has been provided for the rejection of the null form of Hypothesis 4d, viz., that there will be no interaction between scoring procedure and text and deletion rate to affect the relationship between cloze and ELBA.

Because of the small numbers of subjects for each cloze test, and because the two-way analysis of variance (Table 7.1, a-e) of Chapter 7, had shown no significant differences between deletion rates, despite the existence of significant interactions between text and deletion rates, it was considered justifiable to group together all four cloze tests, of differing deletion frequency, on the same text to provide a composite cloze test for each text. The results of the correlation of this grouping with ELBA are presented in Table 8.7.

TABLE 8.7

Pearson Product Moment correlations of cloze scores with ELBA and dictation by test.

a) Difficult Text

	ELBA										Dictation	
	1	2	3	4	Pt1	5	6	7	Pt2	Total	I	II
Cloze												
Exact	.36	.49	.25	.55	.51	.71	.57	.54	.69	.66	.50	.54
SEMAC	.50	.57	.30	.67	.65	.79	.73	.63	.82	.80	.58	.59
GRCO	.37	.49	.35	.50	.51	.72	.61	.50	.71	.66	.52	.51
IDFC	.41	.53	.34	.57	.56	.74	.64	.50	.73	.71	.52	.52
ACFC	.39	.50	.33	.56	.55	.71	.64	.50	.72	.69	.51	.51

(ELBA: Cloze, n = 81; Dict I: Cloze, n = 87; Dict II: Cloze, n = 88. $p < .001$)

b) Medium Text

											Dictation	
	1	2	3	4	Pt1	5	6	7	Pt2	Total	I	II
Cloze												
Exact	.49	.36	.36	.56	.63	.58	.55	.51	.64	.67	.58	.58
SEMAC	.57	.47	.43	.70	.76	.65	.63	.59	.72	.78	.69	.69
GRCO	.59	.36	.37	.66	.73	.62	.61	.55	.69	.75	.62	.66
IDFC	.42	.33	.44	.52	.57	.54	.49	.49	.58	.62	.51	.56
ACFC	.43	.29	.38	.52	.57	.53	.48	.50	.58	.61	.51	.56

(ELBA: Cloze, n = 91; Dictation: Cloze, n = 95 $p < .001$)

c) Easy Text

											Dictation	
	1	2	3	4	Pt1	5	6	7	Pt2	Total	I	II
Cloze												
Exact	.47	.47	.29	.63	.58	.58	.55	.47	.59	.61	.70	.58
SEMAC	.55	.53	.45	.70	.68	.69	.61	.58	.69	.71	.76	.67
GRCO	.51	.44	.42	.65	.62	.66	.59	.57	.67	.67	.74	.66
IDFC	.43	.43	.35	.57	.54	.54	.46	.47	.54	.56	.64	.58
ACFC	.43	.39	.35	.58	.54	.56	.48	.49	.56	.57	.69	.62

(ELBA: Cloze, n = 92 Dictation: Cloze, n = 93 $p < .001$)

If one compares these composite tests, the apparent differences between texts are considerably reduced. In particular, the difficult text is no longer pre-eminent as a predictor of performance on ELBA. In fact, there now seems to be little difference between the difficult and medium texts in terms of ability to predict English-as-a-foreign-language proficiency (exact: .66 and .67; SEMAC: .80 and .78). However, the grammatical scoring procedures do not agree with this pattern - the grammatically-correct-response scoring procedure (GRCO) shows the medium text to be by far the best predictor, whereas both form class procedures (IDFC and ACFC) give the advantage to the difficult text. There is, nonetheless, overall general agreement among scoring procedures that the easy text is a worse predictor of English-as-a-foreign-language proficiency than either of the other two texts.

On the whole, this agreement also holds when the individual test correlations are examined. Table 8.8 is a recast of part of Table 8.7, to enable an easier comparison of texts as predictors of individual ELBA tests.

TABLE 8.8

Comparison of cloze texts as predictors of individual ELBA tests.

Exact word and any-acceptable scoring procedures only.

1) Exact word score

	Cloze Text		
ELBA Test	Difficult	Medium	Easy
1	.36	.49	.47
2	.49	.36	.47
3	.25	.36	.29
4	.55	.56	.63
5	.71	.58	.58
6	.57	.55	.55
7	.54	.51	.47
Pt1	.51	.63	.58
Pt2	.69	.64	.59
Total	.66	.67	.61

2) Any-acceptable-word score

	Cloze Test		
ELBA Test	Difficult	Medium	Easy
1	.50	.57	.55
2	.57	.47	.53
3	.30	.43	.45
4	.67	.70	.70
5	.79	.65	.69
6	.73	.63	.61
7	.63	.59	.58
Pt1	.65	.76	.68
Pt2	.82	.72	.69
Total	.80	.78	.71

This table shows that cloze on an easy text is generally the poorest predictor of performance on a reading comprehension test (7), a vocabulary test (6), a grammar test (5), and of performance on a battery of listening tests (subtotal, Part 1), and a battery of reading tests (subtotal, Part 2). However, this does not apply to the individual listening tests, and the easy text is at least as good a predictor of sound recognition ability, intonation recognition and listening comprehension as the medium text.

The lack of difference between medium and difficult texts as predictors of overall ELBA scores is not maintained for the prediction of individual test scores. For both SEMAC and Exact scoring procedures, the medium text is a better predictor of performance on Part 1 (listening) of ELBA, whereas the difficult text is a better predictor of performance on Part 2 (reading). Furthermore, the difficult text is a better predictor than the other two texts of performance on the reading tests of Grammar (5), Vocabulary (6) and Reading Comprehension (7) - again, for both the SEMAC and the Exact procedures. Since, although there is little difference between the difficult and medium texts as predictors of overall English-as-a-foreign-language ability, there is a marked advantage of the difficult text as a predictor of relevant ELBA tests (relevant, that is, at least superficially to the cloze test itself), the evidence confirms that provided by an inspection of the individual cloze tests to justify a rejection of the null form of Hypothesis 4b.

In order to see whether the individual tests of the ELBA battery relate differently, i.e., in a different order, to each cloze

text, the ELBA tests were ranked according to the size of their correlation with the cloze texts, and the results are presented in Table 8.9.

TABLE 8.9

Rank orders of correlations of ELBA individual tests with cloze text. All scoring procedures.

1) Difficult Text

Scoring Procedure	ELBA Test 1	2	3	4	5	6	7
Exact	6	5	7	3	1	2	4
SEMAC	6	5	7	3	1	2	4
GRCO	6	5	7	3=	1	2	3=
IDFC	6	4	7	3	1	2	5
ACFC	6	4=	7	3	1	2	4=

2) Medium Text

Scoring Procedure	ELBA Test 1	2	3	4	5	6	7
Exact	5	6=	6=	2	1	3	4
SEMAC	5	6	7	1	2	3	4
GRCO	4	7	6	1	2	3	5
IDFC	6	7	5	2	1	3=	3=
ACFC	5	7	6	2	1	4	3

3) Easy Text

Scoring Procedure	ELBA Test 1	2	3	4	5	6	7
Exact	4=	4=	7	1	2	3	4=
SEMAC	5	6	7	1	2	3	4
GRCO	5	6	7	2	1	3	4
IDFC	5=	5=	7	1	2	4	3
ACFC	5	6	7	1	2	4	3

Tests 1, 2 and 3 (Sound Recognition, Intonation, and Stress) all clearly show relatively low correlations with cloze, regardless of the text used, whereas Tests 4, 5, 6 and 7 (Listening Comprehension, Grammar, Vocabulary, and Reading Comprehension) all show a closer relationship, again regardless of text. The striking thing about this table, apart from the agreement amongst scoring procedures which will be examined in the next section, is the high agreement among texts as to the relative importance of the different individual ELBA tests. The difference between the texts results from a change in the order of importance of the first three tests only, viz., Tests 4, 5 and 6. On the difficult text, grammatical abilities seem to be most important, followed by knowledge of vocabulary, and then listening comprehension, which, as we have seen, is not entirely listening comprehension. The medium text differs only in that, whilst still giving prominence to grammar, the importance of vocabulary is somewhat reduced, and that of listening comprehension is increased. The easy text continues this trend of increasing the importance of listening comprehension, whilst downgrading both grammar and vocabulary to second and third places respectively. In other words, the difficult text seems to measure language elements - structure and lexis - more than comprehension, whilst the other two texts give more importance to comprehension. Interestingly, however, in all cases the importance of reading comprehension (Test 7) remains fairly small, and in fact decreases in absolute terms (Table 8.8) as the text becomes easier.

The foregoing appears to provide adequate justification for the rejection of the null form of Hypothesis 4b, and the contingent

conclusion that differences do exist among texts used in cloze tests as measures of proficiency in English as a foreign language.

8.2.3 Scoring procedures

As was noted in the previous section, the evidence from Table 8.9 showed that the correlation coefficients for each scoring procedure with the ELBA tests were ranked in roughly the same order for each procedure on any one text. On the difficult text, the only difference in relative rank ordering between scoring procedures was the position of Test 7 - Reading Comprehension. This varying position of the reading comprehension test can also be seen in the medium and easy texts, together with a certain disagreement over the relative positioning of Test 4 (Listening Comprehension) and Test 5 (Grammar). The exact and any-acceptable-word scoring procedures largely agree on the relative size of coefficients for all texts, but the grammatical scoring procedures show some disagreement with each other, and with the other two scoring procedures. As might be expected the form class scores correlate highest with the ELBA grammar test, for both difficult and medium texts, but not, curiously enough, for the easy text. Similarly, the grammatically correct (GRCO) procedure does not always correlate highest with the grammar test - on the medium text, it correlates highest with Test 4 (Listening Comprehension). These few differences, however, are relatively unimportant and do not detract from the striking overall agreement among scoring procedures.

Although the rank order of the coefficients is roughly the same for each scoring procedure, the size of the coefficients varies from procedure to procedure. It is possible to recast Table 8.7, by

ranking the correlations for any one ELBA test on any one cloze test, over the five scoring procedures, to indicate which scoring procedure correlates highest, among scoring procedures, with any given ELBA test, subtotal or total. This is done in Table 8.10.

T A B L E 8 . 1 0

Rank order of correlations of scoring procedures with ELBA tests, grouped by cloze text.

1) Difficult Text

	ELBA									
Score	1	2	3	4	5	6	7	Pt1	Pt2	Total
Exact	5	4=	5	4	4=	5	2	4=	5	4=
SEMAC	1	1	4	1	1	1	1	1	1	1
GRCO	4	4=	1	5	3	4	3=	4=	4	4=
IDFC	2	2	2	2	2	2=	3=	2	2	2
ACFC	3	3	3	3	4=	2=	3=	3	3	3

2) Medium Text

	ELBA									
Score	1	2	3	4	5	6	7	Pt1	Pt2	Total
Exact	3	2=	5	3	3	3	3	3	3	3
SEMAC	2	1	2	1	1	1	1	1	1	1
GRCO	1	2=	4	2	2	2	2	2	2	2
IDFC	5	4	1	4=	4	4	5	4=	4=	4
ACFC	4	5	3	4=	5	5	4	4=	4=	5

3) Easy Text

	ELBA									
Score	1	2	3	4	5	6	7	Pt1	Pt2	Total
Exact	3	2	5	3	3	3	4=	3	3	3
SEMAC	1	1	1	1	1	1	1	1	1	1
GRCO	2	3	2	2	2	2	2	2	2	2
IDFC	4=	4	3=	5	5	5	4=	4=	5	5
ACFC	4=	5	3=	4	4	4	3	4=	4	4

If we begin by looking at the correlation with the ELBA Total, it is quite clear that the any-acceptable-word (SEMAC) procedure always shows the closest relationship, regardless of text used. The easy and medium texts agree that the grammatically correct (GRCO) procedure is the next best predictor of overall ELBA scores, with the exact word scoring procedure coming a poor third. Either form class procedures (IDFC or ACFC) are about as good as the other in predicting ELBA scores, and worse than the other procedures. Nevertheless, the correlation is still quite reasonable, at .56 to .62, for even the poorest predictors. On the difficult text, the situation is somewhat different, since both the exact word and GRCO correlate lower than either of the two form class scoring procedures. This is clearly because, contrary to the general trend (namely, that correlations of cloze with the ELBA Total increase with increasing text difficulty - SEMAC, IDFC, ACFC all do this), both the exact and GRCO procedures have a lower correlation with the difficult text than with the medium text.

This evidence is sufficient, however, to enable the rejection of the null form of Hypothesis 4c, and to support the contention that different scoring procedures do indeed relate differently to measures of English-as-a-foreign-language proficiency, and that the best predictor of such proficiency is not the exact word method, but the scoring procedure which allows any semantically acceptable word as correct.

This conclusion is supported by the correlations of each scoring procedure with individual ELBA tests, and their subtotals. (Table 8.10). The any-acceptable-word procedure is always the best predictor of the part totals - listening comprehension and reading compre-

hension - with the GRCO procedure in second place for the easy and medium texts, and the IDFC in second place on the difficult text. The exact word method is never higher than third place, whilst for Part 2 on the difficult text, it is in fact the worst predictor. (Interestingly, despite this poor prediction, the exact word method is unlike the SEMAC and GRCO procedures, but similar to the form class procedures, in that it correlates consistently higher with the Part 2 (Reading) subtotal than with the Part 1 (Listening) subtotal, regardless of the text used. The SEMAC and GRCO procedures differ in that on the medium text, the Part 1 subtotal correlates higher than that of Part 2. The implication of this will be examined later, in conjunction with an examination of the relative importance of dictation.

With only three unimportant exceptions (Test 3 (Stress) on the difficult and medium texts, and Test 1 (Sound Recognition) on the medium text), the SEMAC always correlates higher with the ELBA tests than any other scoring procedure. It always, without exception, correlates higher than the exact word procedure, whatever the text or ELBA test. As noted above with the ELBA totals and part totals, on the difficult text the form class scores correlate higher with the individual ELBA tests than the exact word and GRCO procedures, but on the other two texts, they are virtually always the worst predictors of the individual tests. Regardless of text, the GRCO is, with only three exceptions, a better predictor than the exact word procedure, and, on the easy and medium texts, is the second best predictor after the SEMAC. Finally, Table 8.10 confirms the evidence from Table 8.9 that no scoring procedure consistently correlates more with certain subtests only than any other scoring procedure does.

In other words, it is not the case that, for example, the GRCO shows a closer relationship to the Grammar Test (Test 5), whilst the SEMAC is more related to vocabulary (Test 6), or the exact word procedure more related to reading comprehension (Test 7). The only possible exception to this is that the exact word method appears to be unusually unrelated to the ability to detect sentence stress (Test 3).

As would be expected, due to the lower number of subjects taking each individual cloze test, and the resulting lower reliability, the correlations based on the results of each cloze test are more varied than those based on data grouped over one text. The great variability of the correlations with individual ELBA tests is, of course, further evidence for the rejection of the null form of Hypothesis 4d, in that the interaction of scoring procedure, deletion rate and text clearly results in different correlations with tests of proficiency in English as a foreign language.

Nevertheless, if one recasts the data from Table 8.4 in the form of rank orders of the coefficients of each scoring procedure with each ELBA test, for each cloze test (Table 8.11), it is possible to discern the same trends as those already noted. In particular, whenever the coefficients are significant, the any-acceptable-word procedure usually shows the highest correlation with individual ELBA tests, subtotals or total. Only once does the exact word procedure correlate higher with the ELBA Total than any other scoring procedure, even the SEMAC (Test E08). On the difficult text, it is usually in second place, behind the SEMAC (except for Test D10, where it lies at joint bottom); on the other texts it is lower, usually correlating less well than the

GRCO procedure. On three out of the twelve tests, in fact, it correlates worst of all with the ELBA Total (D10, M10, E12). The range of coefficients for the exact method is from .51 to .86, which is greater than that for the SEMAC (.67 to .88).

The GRCO correlates highest with the ELBA Total on two tests (M10, E10), but also correlates lower than the other scoring procedures on all four difficult-text tests. The form class scores tend to correlate worst. On six of the twelve tests they are lower than the other three scoring procedures, but occasionally they are higher; all the difficult-text tests, and E12, show higher correlations when scored by the form class scores than when scored by GRCO. On Tests D10, M10 and E12 they correlate with ELBA higher than the exact word procedure does. However, on only one test - D10 - is one form class score (IDFC) as high as the SEMAC.

8.3 Summary of findings

8.3.1 ELBA alone

1) On the ELBA test, Tests 4, 5, 6 and 7 (Listening Comprehension, Grammar, Vocabulary, and Reading Comprehension) were closely related. In particular, Tests 4, 5 and 6 were the major predictors of the total score. Test 7 was less important, and Tests 1, 2 and 3 even less so.

2) A factor analysis of the ELBA produced two alternative solutions:

a) ELBA is unifactorial, meaning that it tests general English-as-a-foreign-language proficiency.

b) Broadly, three factors are involved in the ELBA Battery - general proficiency; inferential abilities, unrelated to

the mode of presentation; and segmental/discrete listening abilities.

3) ELBA is best seen, not as a test of listening and reading, as the division into Parts 1 and 2 implies, but as a test of "core proficiency" and "inferential abilities", or, to put it another way, of ability to handle sentences and ability to handle text.

8.3.2 Hypothesis 4

The null forms of Hypothesis 4 were rejected, and it was concluded that differences exist between deletion rates, between texts, and between scoring methods as measures of proficiency in English as a foreign language. It was also concluded that there was an interaction between these variables which affects the ability of a cloze test to predict English-as-a-foreign-language proficiency.

8.3.3 Deletion rate effect

1) The differing correlations of deletion rates with ELBA are not predictable. It is thus not possible to conclude that a more frequent deletion frequency is a better measure of English-as-a-foreign-language proficiency.

2) There is no patterning of relationships of certain deletion rates with certain ELBA tests. Thus it is not possible to conclude that a more frequent deletion gives a different measure of English-as-a-foreign-language proficiency.

8.3.4 Text effect

1) If deletion rate differences are taken into account, the difficult text appears to correlate highest with a measure of English as a foreign language.

2) If deletion rate differences are ignored, the difference between texts as measures of such proficiency is reduced.

3) The easy text is generally the poorest predictor of ELBA.

4) On the whole, no one text measures more of a skill than any other text does, but two tendencies are discernible:

a) As the text becomes easier, grammar and vocabulary abilities are less associated with performance on the cloze.

b) The difficult text tends to be the best predictor of "core proficiency" (Tests 5 and 6). Perhaps the difficult text measures language elements whereas the easier texts measure more comprehension.

8.3.5 Scoring procedure effect

1) The any-acceptable-word scoring method is always the best predictor of overall English-as-a-foreign-language proficiency.

2) The grammatically-correct procedure is second best, and better than the exact word method.

3) Both the SEMAC and the GRCO procedures are usually better predictors of individual tests than the exact word method.

4) The exact word method tends to emphasize the differences between deletion rates and between texts.

5) The SEMAC tends to reduce the differences between deletion rates and between texts.

6) Except when differences between deletion rates are ignored, the SEMAC increases the difficult text's superiority over the medium text as a predictor of core proficiency.

7) No scoring procedure measures more of one skill than another skill. The SEMAC is not a test of vocabulary, and the GRCO is not a test of grammar.

8.3.6 "Best buy"

The "best buy" as a measure of proficiency in English as a foreign language seems to be a difficult passage, scored by the any-acceptable-word procedure.

8.3.7 What cloze tests

In general, Tests 4, 5 and 6 (Listening Comprehension, Grammar, and Vocabulary) are more closely related to cloze than the other tests are. Tests 2 and 3 (Intonation and Stress) are the least associated. Test 7 (Reading Comprehension) is of relatively little importance in the cloze. This would seem to mean that the cloze is more a measure of "core proficiency" than of reading comprehension.

8.4 Results of the dictation tests

There were 275 students tested on Dictation I and cloze; 276 on Dictation II and cloze. Of these students, 197 also took ELBA.

The two dictations were scored, as mentioned in Chapter 5, according to Oller's system of counting one error per word, unless it was a spelling error, and the difficult dictation was also scored according to Fountain"s system of counting only key words. The results are tabulated in Table 8.12. Dictation I had a mean of 90%, and a standard deviation of 9.2%; Dictation II had a mean of 75%, and a standard deviation of 17.7%. When scored according to Fountain's system (key words only) the mean of Dictation II was 76.9%, and the standard deviation was 18.2%.

8.4.1 Comments on the scoring procedures

Although spelling errors are usually obvious, there were occasions when an error could have been either graphological or, more seriously, semantic. Thus "hire education" for "higher education" could merely reveal the subject's inability to spell, or it may indicate a genuine lack of comprehension. He does seem, however, to have heard the sounds correctly. For this reason "hire education" was counted a correct version of "higher education". Another frequent problem was the occurrence of "were" for "where". In the scoring, this was assumed to be a spelling error, but it could just as validly have been considered a syntactic error. Also classed as spelling problems were errors like "machin" for "machine", although the apparent lack of knowledge of phoneme-grapheme correspondence - /i:n/ = "-ine", /in/ = "in" - seems to be of a different order from errors like "aplied" for "applied". The first point, then is that to exclude all spelling errors from an error count is not as easy as it sounds.

Secondly, the question of weighting errors arises frequently, since it appears that errors like "jops" for "jobs", or "partry" for "partly" are of a different order, i.e., show a different degree of lack of comprehension, to errors like "sentence" for "centres", or "this grief" for "this drift". Indeed, one could argue that "jops" and "partry" do not indicate lack of comprehension. This problem was not resolved here - its solution is probably a study in its own right. It must merely be noted that the dictation scores referred to do not distinguish between major and minor errors. Further, one should also note that intrusions (e.g. "into the possession" for "into possession") as

well as inversions of word order, were classed as incorrect by the first scoring system.

Thirdly, under the first scoring system, where any correct word results, in effect, in a plus point, even the utmost gibberish gets a score of between 25% and 30%. This was relatively common in Dictation II, where it can be assumed that any script with a score below about 44% made very little sense. Subjects occasionally, in the second half of the second dictation, wrote only the function words, yet that, in the second half alone, would give them a score of around 24%.

Fourthly, the second scoring system, as used by Fountain (i.e., mark only key lexical words) is intended to avoid this problem. However, the obvious question is whether or not this scoring scheme tests comprehension. If one "correctly" recognises the nouns and verbs, has one necessarily understood the content? Is

"The shift to the towns has decreased the necessity for these facilities."

the same as

"A shift in town is decreased these necessity from his facilities."?

And yet the key words (underlined) are correct in both cases. Is

"the fact obviously been calculated approximately"

to be given two points, as a version of

"despite the obvious need to calculate approximately"?

Fifthly, morphological errors, such as omission or addition of final -s, -es, -d, -ed, were ignored in the Fountain scheme as used here. How does this affect the scoring of "the used machines on farms" for "the use of machines on farms"? Clearly the meaning is different, so this

type of error was classed as syntactico-semantic rather than morphological, and thus counted.

These remarks are intended to emphasize the fact that some scoring decisions are necessarily arbitrary, or judgemental, and that any scoring scheme has pitfalls.

8.4.2 Relationship between Fountain and ordinary scoring scheme, Dictation II

Because of the experimental design, it was possible to compare two different procedures for scoring the difficult dictation, to see how they differed as tests and how they related to external criteria, in order to throw some light on what dictation tests.

The ordinary (Oller) procedure resulted in a mean which was 75% of the maximum possible (Table 8.12). The key word procedure (Fountain) had almost exactly the same result (mean = 77%). In both cases, the standard deviation was 24% of the mean. However, the distribution of the Fountain procedure was more negatively skewed than that of the ordinary procedure, resulting in greater top-heaviness. Whereas 21% of the subjects scored over 90% on the ordinary procedure, 31% scored over 90% on the key word procedure. In other words, the key word procedure is less able to discriminate amongst the better students than is the ordinary procedure (both are, of course, better than the easy dictation, which resulted in 65% of subjects getting over 90%). To some extent, this finding is to be expected, since there are far fewer items scored in the key words procedure than in the ordinary procedure (90 versus 182). And, it may be that what one in effect is saying, is that the good students understand most of even a very difficult dictation - in

that they hear, comprehend and reproduce the lexical items - but that they may have some difficulty with, or make some errors in, the syntax of the text, possibly because they do not understand it, or because their productive abilities are still somewhat faulty, or they do not notice or hear some of the function words. After all, these function words are to some extent redundant, and usually "spoken with less clarity" (more reduced vowels, more elision, omissions, etc.) than the lexical items.

To what extent, however, can one be said to have understood a passage if one has understood the nouns, verbs, etc., in it? It is true that the ordinary scheme discriminates among the better students somewhat more satisfactorily than does the Fountain scheme, but the question is - to what extent is this discrimination valid, or to what extent is it a spurious discrimination?

If the discrimination is a true one - i.e., it reflects real differences in ability - then one might expect the two procedures to relate differently to other measures of language proficiency. Perhaps the ordinary scheme relates more closely to grammatical proficiency than does the key word scheme, or perhaps the key word scheme is not so closely related to listening comprehension, especially overall listening ability.

However, as Table 8.13 shows, this is not the case.

TABLE 8.13

Correlations of dictation tests with ELBA

(Spearman rho rank correlation)

ELBA	1	2	3	4	Pt1	5	6	7	Pt2	Total
Dictation I (easy)	.74	.40	.35	.72	.79	.70	.72	.54	.75	.81
Dictation II (difficult)	.66	.35	.27	.68	.71	.65	.70	.57	.72	.75
Dictation II (key word)	.65	.35	.25	.68	.70	.64	.74	.55	.73	.75

n = 197

In fact, the correlations of the two scoring procedures with the ELBA total score are identical (.75) and the correlations with the part scores on ELBA show little or no difference. The sole exception to this is the correlation with the Vocabulary subtest, where the key word scheme correlated .04 higher than the ordinary scheme, and this difference seems surprisingly small. Moreover, this agreement between scoring procedures as to external criteria is the same for all the sub-populations, as Table 8.15 shows. In fact, it seems that the two procedures are measuring the same thing. This is confirmed by the high intercorrelation between the two (.97). The only difference seems to be that the ordinary scheme is somewhat better at discriminating among the better students. What is the implication of this? If both schemes are measuring the same thing, yet are at least superficially different, what are they measuring?

The dictation task is the same under both conditions, of course, since the two procedures are merely scoring procedures. So the

tests measure the same thing, the only question being whether one of the two procedures samples the task more accurately or efficiently than the other. As far as accuracy is concerned, whatever ability it is that a dictation test taps seems to be tapped as accurately by one procedure as by the other since the correlations with external criteria are essentially the same, and since the two schemes intercorrelate so highly.

As far as efficiency is concerned, however, the ordinary scoring scheme seems to have the edge over the key word scheme, because of the more normal distribution associated with it. The key word scheme is not as sensitive to small but real differences in performance among the better students. Whether this lack of sensitivity is due to the quality of the sample, or its quantity, is the problem. To put this another way, what differences there are between good students on a dictation test may lie in their ability to reconstruct the syntax of the text, rather than to understand the content, and thus the ordinary method would be better because it is qualitatively different from the key words scheme, which samples only one area of language. The alternative explanation is that the ordinary scoring scheme fares better simply because it has more items. The key word scheme, on the other hand, only looks at some of the evidence, although the sample it takes is valid. If the schemes were qualitatively different, then one would conclude that there are small but important differences between students even at an advanced level, but these differences lie not in the ability to hear or understand lexical items, but in the ability to relate these items to the overall message structure of the text. But if it were true that one scoring procedure were able to show up this difference better than

another, one would expect validating correlations to vary. The fact that they do not is evidence for the contention that the difference between the procedures is purely quantitative.

The point, then, is that dictation remains, if you like, an integrative task; although one of the scoring procedures apparently ignores this by concentrating on one aspect of the task - the ability to hear and understand the content - it is no more or less integrative than the ordinary scheme. The advantage of the latter scheme is that it samples more of the same thing and so is capable of making more discriminations.

The conclusions to be drawn from all this are:

1) There is no qualitative difference between scoring schemes.
2) (a theoretical point) Measuring the subject's ability to identify nouns, verbs and the like correctly gives the same results as including a measure of the syntax as well. This does not mean that doing dictation is essentially a matter of identifying correctly all the lexical items.
3) The reason for the correspondence between the two schemes is that the key word scheme provides a good sample of the success of individuals in the dictation task. The reason the ordinary scheme provides better discrimination is because it provides more, not better, evidence of the performance on the task.
4) Since the ordinary scheme is more efficient, the key word scheme will be ignored in ensuing discussions of dictation.

Although it is conceivable that people doing a dictation might write down the lexical items correctly, and the syntax incorrectly,

in fact, they do not. The misunderstandings and incorrect representations are reflected both in the syntax and in the lexis. Consider the errors of comprehension and production in the reproduction of:

"But the total explanation of this drift is more involved."

as found in the sentence (which actually occurred in a subject's paper):

"But the title explanation of this script is more enjoyed".

The original sentence has four key words, of which the subject reproduced only one correctly. Yet the syntax is more or less the same. In fact, if garbage is to be produced, it is as likely to occur in the lexis as in the syntax, and the key word scheme is capable of registering the fact that it has occurred.

8.4.3 The nature of the errors

It has been suggested that some insight into what dictation might be measuring may be gained from a study of errors made in dictation, and typical errors were quoted to reinforce Oller's point that a process of analysis by synthesis is involved. Although an error analysis is beyond the scope of the study, a rapid perusal of the data from the two dictations may give some useful insights. (For samples of errors, see Appendix F.)

Creative errors of the type mentioned by Oller do occur fairly frequently in the difficult text, but not in the easy one (where, indeed, relatively few errors of any sort occur). These errors indicate that when the subject has not understood the phrase, he guesses at its meaning based on one or two sound sequences he has heard. (Put differently, he interprets the sound sequences he has heard differently from the normal interpretation.) Frequently these guesses or interpretations do not fit

into the context of previous and following dictated chunks, and yet they make sense in isolation. "That was why he'd got the place for a song" becomes "That was the place that God placed for a song"; "up the river" becomes "after Riva"; "We had the last piece" becomes "We had a large beach"; "in essential services" becomes "in central surfaces"; "But the total explanation of this drift is more involved" becomes "But the title explanation of this script is more enjoyed".

In most of these examples, some words from the text are repeated verbatim, and it appears that only scraps of the remainder have been heard. These scraps are then interpreted in the light of the understood portions. The process is perhaps shown more clearly in clauses where only one or two words are mistaken, and where the substituted words have sounds in common with the original words:

"The shift to the towns" becomes "The ships to the towns" (/ʃɪ /);

"the growth of industry" becomes "the growth over the streeet" (/v/, /stri/);

"the main centres offer" becomes "the man sent his offer";

"the most remote" becomes "the most remarked".

Rivers' opinion (Rivers, 1968) that students "don't pay attention to the way segments fit into the passage" may be valid, but it is certainly not true to say, as she does, that they do not pay attention to the meaning of what they write. Subjects are desperately trying to make sense of what they hear, hence the fairly common creative errors. Indeed, their desire to make sense of each chunk may lead them to produce strings that do not fit into the passage as a whole. Rarely is garbage produced <u>within</u> the clause, whatever happens outside it; it is noticeable that

these creative errors are always syntactically correct.

A second, less common but equally interesting category is exemplified by the following:

"would otherwise be engaged in agriculture"

becomes

"would otherwise be engaged in the farm".

This type could be called a lexical error, where the overall sense of the chunk, and its meaning within the context, are retained, but realised by a different surface item. Further examples from the data are:

"the institutions providing higher education"

becomes

"the institutions offering higher education";

"a significant proportion of the population"

becomes

"a significant proportion of the people";

"are generally located in the towns"

becomes

"are generally placed in the towns";

"despite the obvious need"

becomes

"in spite of the clear need";

"necessity"

becomes

"need".

Two explanations seem possible. The first is that these are simply more successful guesses than the previous examples. This

possibility is weakened in many cases, by the lack of phonetic similarity of the substitution to the original. The second explanation is that the correct word was indeed heard and understood, but that what was remembered were the semantic features, rather than the surface phonological forms. This could be further evidence for active participation in the listening process, since some kind of transformation is clearly involved.

This notion of active participation, where the listener extracts, as it were, "deep" features from what he is processing, retains these features in memory, and then actively produces the chunk (in this case, in writing) from these features as if the chunk were a novel utterance of his own, is supported by errors like "left" becoming "leaved", where the listener could be said to have extracted the lexeme LEAVE, plus the feature PAST TENSE, and then on reproducing, to have utilised his normal language production rules to produce the deviant "leaved".

This explanation would also account for the following two types of error found in the data, namely, 1) word order errors, where elements are transposed in reproduction (which confounds Lado's statement that word order presents no problems to the dictatee), and 2) phonological errors of the type "jops" for "jobs", or "partly", "during", "correctly" becoming "partry", "duling", "collectly". (This latter type of error is infrequent, as one would expect, since it is almost exclusively an error of <u>speech</u> production.)

One should beware of inferring too much about the listening process from examples of when the comprehension process seems to have

broken down. It is possible that the process can operate this way, especially when under stress, but that under normal conditions, it does not do so. The problem, of course, with using dictation errors as evidence of listening comprehension processes in action, or to see whether dictation measures listening comprehension, is that the errors may be due to production, and not to comprehension. We do not know whether the signal is only partially received, and thus faultily reproduced; whether the signal was received perfectly, and only reproduced faultily; or, indeed, whether, as suggested, the signal is received whole, the significant information extracted and stored, and then regenerated for production.

8.4.4 Dictation as a test

"It is a whole lot harder to construct a good multiple-choice vocabulary test than it is to construct a good dictation" (Oller and Streiff, 1975). Unfortunately, Oller does not indicate what a good dictation test is, nor how one is to tell that he has constructed one. If by "good" he means efficient, then the evidence from this study indicates that he is almost certainly wrong. Several people in the discussion of Oller's paper make the point that there is a lot of "dead data" in dictation. The evidence from this study indicates that the amount of dead data depends on the passage used. In Dictation I there were only about a dozen regions in which subjects made errors - the rest was almost entirely error-free, and even the weak students had no trouble. Certainly, even advanced students did make errors, but so few as to provide very little discrimination. Note that 85 students out of 275 had a score of at least 180 (maximum 187) on this dictation, 65%

got scores of 90% and over, and the lowest score was 94, i.e., over 50%. The mean score was 90%. This test, at least, is relatively inefficient. On the difficult passage, Dictation II, errors were more evenly spread through the passage, with a tendency to increase towards the end. Very few errors, however, were noted in the first sentence. Even in this second dictation, the lowest score was 41 out of 182, and some 21% of the population scored 90% and over. The mean was 75% of the total number of items. It would appear that, bearing in mind the number of items in the dictation tests, they are not particularly efficient tests, and, probably, they do not discriminate adequately the fairly advanced students. It is hard to imagine any dictation that would, if Dictation II could not.

Since Oller also referred, earlier in the discussion, to the way a dictation ranks subjects, it is possible that by a good test, he means one that provides a wide spread. In fact, he asserts, "typically, on a 50-point test, the standard deviation is about 15 points", whereas a standard grammar test of the same number of items is said to give a standard deviation of 8 or 9 points. This may be irrelevant: we have seen in our discussion of scoring procedures how arbitrary decisions can be. We have also pointed out that usually, as in this study, no distinction is made between major and minor errors. Thus the spread of dictation scores could be quite unreal, since the difference between dictation scores of 35 and 36 is not necessarily the same as the difference between, say, grammar test scores of 35 and 36. Moreover, the spread achieved with Dictation I is much less than the 30% said to be typical. In fact the standard deviation is a mere 10% of the mean.

Dictation II is considerably better, with a standard deviation 24% of the mean. Yet in both cases, the distribution is negatively skewed, and the standard deviation is increased by the long tail of weak subjects. For these two dictation passages, the spread does not correspond to what it said to be typical for dictation.

In both the dictations, the reliability was admirably high (KR21 = .94 and .97 respectively), but this is very much a function of the number of items used.

8.4.5 The difference between the two dictations

> "It doesn't make a whole lot of difference whether you take a fairly hard passage, a fairly easy one or one somewhere in the middle. The test seems to perform similarly and the correlations you get with external validating criteria are similar." (Oller and Streiff, 1975)

Unfortunately Oller does not provide empirical evidence to back up his remarks. However, the following points emerge from the current study:

1) The difficult dictation (II) is the more efficient, as a test. 65% of the subjects attained a score of at least 90% on the easy dictation, whereas only 21% scored above 90% on the difficult dictation. The easy dictation showed fewer errors made in far fewer places, and therefore contained more "dead data" than the difficult dictation. To this extent, it is not true that the tests perform similarly, and Dictation II is to be preferred.
2) Tables 8.13 and 8.14 throw some light on the question of whether the two dictations rank subjects similarly. Their intercorrelation

(Spearman rank) is .82, which indicates a considerable measure of agreement, but still leaves a not-inconsiderable amount of variance unaccounted for. Wether the two can be said to rank subjects similarly depends on how similar "similar" should be. On the whole, since one would not expect negative or zero correlations between two such similar tests, one might well think that .82 was not high enough.

3) Oller's third point is that correlations with external criteria are similar. Again, the problem is the definition of "similar". Table 8.13 shows that whilst the correlations with ELBA show the same pattern, the coefficients are by no means identical. With the ELBA total score (i.e., some sort of overall proficiency measure) the easy dictation correlates .81, the difficult dictation .75; with the listening subtotal the respective correlations are .79 and .71; and with the reading subtotal .75 and .72. It can be seen that the easy dictation correlates consistently higher with external criteria than does the difficult dictation (Table 8.13). This is also true for all 12 cloze test subgroups of the population; Table 8.15 shows only two of the twelve correlations with ELBA Total to be higher for the difficult dictation than the easy one. The conclusion is that although both dictations show a "similar" pattern of correlations with the subtests, the easy dictation is a consistently better predictor of proficiency on ELBA than the difficult dictation.

4) The correlation of both dictations with the twelve different cloze tests, scored in five different ways, is shown in Table 8.16.

When scored by the exact word method, cloze correlates more highly with the easy dictation than with the difficult dictation, regardless of text or deletion rate differences, with only one minor exception. The coefficients range from .44 to .82 for the easy dictation, and from .30 to .77 for the difficult dictation.

The same relative performance of the two dictations is true when compared with cloze scored for any acceptable word, and cloze scored for any grammatically correct word, whilst the two form class scoring procedures each show ten of the twelve correlations as higher for the easy dictation.

Here, too, then, one must conclude that regardless of text used, of deletion rate, or of scoring procedure, the cloze virtually always correlates more highly with the easy dictation than with the difficult dictation.

Thus it appears that the best dictation to use as a predictor of other proficiency measures is a relatively easy text. Difficult texts heard only once consistently perform worse, although never by much. However, the difficult dictation is a much more efficient test, whilst still being far from ideally efficient, and is presumably to be preferred for testing reasons.

The discussion of a possible explanation of this difference between the two dictations will be postponed until the next section.

8.4.6 What does the dictation test?

In any consideration of what dictation is measuring, the first point to be made is that it is clearly closely related to proficiency in English (.81 easy, .75 difficult, in this study). Moreover,

this is true regardless of which subgroups of the population one samples - the correlations of ELBA Total with Dictation I and Dictation II, for each cloze test group range from .57 to .92, with 13 out of the 24 correlations being over .80 (Table 8.13). There can be no doubt that dictation is a reasonable measure of English proficiency.

The second point to be made is that dictation correlates at least as high with the listening test subtotal of ELBA (.79 easy, .71 difficult) as with the reading test subtotal (.75, .72) and in the case of the easy dictation, more highly. This is somewhat contrary to previous findings, but not greatly. Higher correlations with listening would doubtless have been gained if dictation had not correlated so poorly with the intonation and stress sections of the ELBA listening subtest. It follows from this that dictation relates as closely to listening ability as to reading ability, and so can be seen as a reasonably general measure of English-as-a-foreign-language proficiency. Involved in dictation are not only an ability to hear phonemic distinctions and relate them to the written word, but also the abilities to understand globally, make appropriate responses, recognise grammatical sentences, and identify word meanings. The ability to identify the tonic of a sentence, and the ability to describe the function of intonation seem to be unrelated to the ability to do dictation, however. This is unlikely to be because these two sections are "discrete point" or "non-integrative", since the sound recognition section (Test 1), although clearly a discrete-point test, correlates highly with dictation (.74, .66), as do the traditionally discrete-point sections of Grammar (.70 and .65) and Vocabulary (.72 and .70). In fact, the reading comprehension

test, which is presumably one of the more integrative sections of the ELBA, correlates much lower (.54 and .57).

This, then, is the third point: that dictation is not clearly related more closely to integrative subtests than to discrete-point subtests.

The fourth point is that dictation is less related to reading comprehension, as measured by ELBA, than previous research would lead us to expect. The high correlation with the reading total is due almost entirely to the high correlation with the grammar and vocabulary sections, and not with reading comprehension. In other words, if dictation is related to reading comprehension, this is due to the relationship between dictation and grammar and vocabulary, rather than to higher-order skills.

The inference from this is that dictation measures lower-order skills, and not the higher-order, inferential, discourse-processing skills. This in turn would help to explain why the easy dictation correlates more highly with English-as-a-foreign-language measures: because it does not demand skills of as high an order as those demanded by the difficult dictation.

Against this background, how does the dictation relate to cloze? On the whole, there seems to be a considerable relationship between the two. The correlations with cloze exact range from .44 to .82 (easy dictation) and .30 to .77 (difficult dictation) (Table 8.16). When cloze is scored by the any-acceptable method, the correlations increase, and range from .47 to .89 (easy dictation), and .39 to .81 (difficult dictation). The other scoring methods all correlate at slightly lower levels (from the easy dictation: IDFC, .35 to .79; ACFC,

.35 to .81; GRCO, .28 to .89). The grammatically correct scoring procedure correlates better than the exact word method on the easy and medium texts, but not on the difficult text.

This seems to bear out the previous conjecture that dictation measures low-level linguistic skills. These particular skills are presumably measured better by the GRCO (grammatically correct) method than by the any-acceptable-word, or exact word methods. The difficult cloze text is something of an exception - since it is very difficult, it is likely that grammatical or low-level skills would not be adequate to cope with the text, and so the GRCO method would not be an adequate measure of these lower-level skills on this kind of text.

It is noticeable that of the three cloze texts, the two easier ones correlate much higher with dictation, regardless of scoring method, than the difficult cloze text. This seems to provide further confirmation of the thesis that students who have inadequate lower-order skills cannot cope adequately with the difficult text, whereas they do have the ability to cope with the easy and medium texts. Similarly, with the dictation, those students who have the basic skills can cope with the task, since it does not demand higher-order skills. This, of course, as we have seen, is especially true of the easy dictation.

The final point to be made is that, contrary to previous research, dictation clearly does not always correlate more highly with cloze than with other measure of English-as-a-foreign-language ability - the two highest coefficients obtained from the 12 subgroups of the population were between dictation and ELBA Total (.92 and .90), not between cloze and dictation. Moreover, dictation correlated lower with

some of the cloze tests than with the ELBA Total or subtotals, although certainly the lowest significant coefficients were between dictation and ELBA, Tests 2 and 3.

The difficult dictation in particular was a rather poor predictor of cloze performance (ranging from non-significant to .77, exact word).

Because of the design of the study, the only way to examine dictation's relationship with the ELBA sections and cloze is to take each cloze test subgroup, select from that group those subjects who have scores on dictation and on ELBA, and then compare the coefficients (this is a different selection from those represented in Table 8.16, who are simply those subjects from each cloze group who also took dictation, regardless of whether or not they also took ELBA). Unfortunately, and inevitably, the numbers in each cell are rather small, but the results are tabulated in Table 8.17.

The first thing that is immediately apparent is that dictation virtually always correlates higher with ELBA Total than with the cloze, scored by any method.

Secondly, cloze exact almost always correlates lower with dictation than do the listening and reading test subtotals. And the listening subtotal correlates better with dictation than does the reading subtotal, two out of three times.

Thirdly, the cloze any-acceptable procedure usually correlates higher than the cloze exact with either of the two dictation tests.

Fourthly, the dictation always correlates higher with at least one of the ELBA sections than it does with cloze exact, and often

correlates higher with three or four sections than with cloze. Only on three out of the twenty-four occasions does dictation correlate higher with the cloze any-acceptable than with any ELBA section.

If one ranks the correlations in order of size and then sums the ranks, the order of importance of the ELBA sections and the cloze exact are as follows:

First: Test 4 (Listening)

Joint Second: Test 1 (Sound Recognition) and Test 6 (Vocabulary)

Fourth: Test 5 (Grammar)

Fifth: cloze exact

Sixth: Test 7 (Reading Comprehension)

Joint Seventh: Test 2 (Intonation) and Test 3 (Stress)

In other words, the dictation is more closely related to listening tests, tests of formal linguistic skills and lexis (and "discrete-point" at that) than to tests of reading comprehension and cloze ("integrative" or otherwise). This does not agree with the findings of Oller or Darnell, but is comprehensible if one ignores the integrative - discrete-point dichotomy, and looks instead at the level of skills being tested by the various tests.

8.5 Cloze, ELBA and dictation

In contrast to the previous section, which examined the correlations of dictation with cloze and ELBA, it is proposed in this section to look at the correlation of cloze with dictation and ELBA. In other words, the relationship between cloze and dictation is examined from a slightly different angle. Where the previous section based its conclusions on an examination of a Spearman rho (rank) correlation

matrix, this section will use as the basic data the Pearson Product Moment correlations of Tables 8.4 and 8.7. This will enable comparisons to be made with the results of the next section (which deals with the results of factor analyses of cloze, ELBA and dictation, which are, of course, based on product moment correlations).

It is clear from Table 8.4 that individual cloze tests correlate quite differently with the dictation, and that the cloze is not consistently more related to dictation than to the ELBA. Test D10, exact word scoring procedure, fails to correlate significantly with dictation, yet the same scoring procedure on Test E10 correlates highly (.86 and .80) with the two dictation tests. Even the any-acceptable procedure shows a wide range of correlations - from .38 on Test D10 to .91 on Test E12. Even within any one text group (easy, medium or difficult) there is a lack of consistency. Where the correlations are generally high, say, on the easy text, one test, E08, shows remarkably low correlations with the dictation. However, this lack of consistency is not confined to the absolute values of the coefficients, it also applies to the value of the coefficients relative to the correlations with ELBA. For ease of inspection, the correlation coefficients from Table 8.4 have been ranked in order of importance for each row - i.e., each scoring procedure on every test - and the rank orders are set out in Table 8.18. This table shows that, on the easy text, the cloze usually correlates higher with dictation than with the ELBA, except for Test E08, whose low coefficients were mentioned above. In this case, cloze correlates better with several of the ELBA tests - Tests 4, 5, 6, Part 1, Part 2, and the total - than with the dictation. Similarly, on the medium text, although cloze relates

more closely to the dictation than ELBA on Tests M08 and M10, on Tests M06 and M12 the relationship is one of the lowest, since not only the ELBA Total and subtotals, but also several of the individual ELBA tests correlate better with cloze than does the dictation. Conversely, whereas in general the cloze tests on the difficult text tend to show a relatively low correlation with dictation, giving particular prominence to ELBA Tests 5 and 6, Test D06 shows a relatively higher correlation of cloze with dictation than cloze with ELBA.

As with the correlations of cloze and ELBA, it seems that different cloze deletion rates will result in different correlations with dictation tests, and these coefficients are different both in absolute value and in relative importance. Just as before, however, there is no consistent patterning to the differences between deletion rates, since on the medium text deletion rates 6 and 12 result in lower correlations, whilst on the difficult text deletion rate 6 results in a relatively <u>higher</u> correlation, and on the easy text, it is deletion rate 8 which results in a <u>reduced</u> correlation. Once more, the evidence is that different deletion rates result in different tests.

Of course, some variation in the correlations may well be due to the comparatively small numbers of subjects taking either cloze and ELBA or cloze and dictation on any one test. Also, the inherent unreliability of the cloze test results will tend to emphasize the lack of stability in the patterning of the coefficients. For this reason, and to remove the effect of different deletion rates on the cloze:ELBA and cloze:dictation correlations, the tests were grouped together, as in the previous section (8.2.2) and the correlations of the text groups with ELBA and dictation are seen in Table 8.7.

Before a detailed examination of this table is undertaken, the relationship of different cloze texts and cloze scoring schemes to the dictation will be examined. Table 8.19 presents the rank order of each scoring procedure as it relates to cloze.

TABLE 8.19

Rank order of scoring procedures correlating with dictation, by text.

	Difficult		Medium		Easy	
	Dict I	Dict II	Dict I	Dict II	Dict I	Dict II
Exact	5	2	3	3	3	4=
SEMAC	1	1	1	1	1	1
GRCO	2=	4=	2	2	2	2
IDFC	2=	3	4=	4=	5	4=
ACFC	4	4=	4=	4=	4	3

This table shows clearly that, as with the ELBA, the any-acceptable word scoring procedure always relates more closely to the dictation tests, *difficult or easy, than any other scoring procedure.* The next best scoring procedure to relate to dictation is the grammatically correct procedure, with one exception, viz., on the difficult text when relating to the difficult dictation. In this case, the exact word procedure produces a higher correlation. In general, however, the exact word procedure is only third best in predicting dictation scores.

Table 8.20 compares the different cloze texts as predictors of the two dictation scores.

T A B L E 8 . 2 0

Correlation of each cloze text with dictation tests, and ranks. (Pearson Product Moment):

Dictation I (Easy)

	Easy	Medium	Difficult	(Rank) E	M	D
Exact	.70	.58	.50	1	2	3
SEMAC	.76	.69	.58	1	2	3
GRCO	.74	.62	.52	1	2	3
IDFC	.64	.51	.52	1	3	2
ACFC	.69	.51	.51	1	2=	2=

Dictation II (Difficult)

	Easy	Medium	Difficult	(Rank) E	M	D
Exact	.58	.58	.54	1=	1=	3
SEMAC	.67	.69	.59	2	1	3
GRCO	.66	.66	.51	1=	1=	3
IDFC	.58	.56	.52	1	2	3
ACFC	.62	.56	.51	1	2	3

The superiority of the easy text is quite marked for the easy dictation and still present, though less so, for the difficult dictation. The difficult text is usually the worst predictor of dictation scores. Interestingly, the text used for the cloze test seems to have an effect on which dictation the cloze score will best relate to.

When a easy text is used, cloze relates much more closely to the easy dictation than to the difficult one; when a medium text is used, there is little difference between the two dictations in their relationship with the cloze; but when a difficult text is used, the cloze tends to correlate higher with the difficult dictation than with the easy dictation. The best correlation overall, however, is clearly between the easy cloze text, and the easy dictation.

In order to compare the relative importance of dictation and the ELBA tests in correlating with cloze, the rank orders of the coefficients for any one scoring procedure on any one text (from Table 8.7) are set out in Table 8.21. This table supplements the evidence from Table 8.20, which showed the absolute values of the cloze:dictation coefficients, by showing the relative importance of the correlation of cloze with dictation.

TABLE 8.21

Rank order of correlations of cloze with ELBA and dictation, by text group.

	ELBA										Dictation	
	1	2	3	4	5	6	7	Pt1	Pt2	Total	Easy	Diff.
Difficult Text												
Exact	11	10	12	5	1	4	6=	8	2	3	9	6=
SEMAC	11	10	12	5	3	4	7	6	1	2	9	8
GRCO	11	10	12	8=	1	4	8=	6=	2	3	5	6=
IDFC	11	7	12	5	1	4	10	6	2	3	8=	8=
ACFC	11	9=	12	5	2	4	9=	6	1	3	7=	7=
Medium Text												
Exact	10	11=	11=	7	4=	8	9	3	2	1	4=	4=
SEMAC	10	11	12	4	7	8	9	2	3	1	5=	5=
GRCO	9	12	11	4=	6=	8	10	2	3	1	6=	4=
IDFC	11	12	10	6	5	8=	8=	3	2	1	7	4
ACFC	10	12	11	6	5	9	8	3	2	1	7	4
Easy Text												
Exact	9=	9=	12	2	5=	8	9=	5=	4	3	1	5=
SEMAC	10	11	12	3	4=	8	9	6	4=	2	1	7
GRCO	10	11	12	6	4=	8	9	7	2=	2=	1	4=
IDFC	10=	10=	12	3	5=	9	8	5=	5=	4	1	2
ACFC	10	11	12	3	5=	9	8	7	5=	4	1	2

The table shows that as the cloze text becomes easier, the cloze correlates more highly with dictation than with ELBA, so that on the easy

text, the cloze correlates higher with the easy dictation than with any other test. However, this does not appear to be true for the difficult dictation, which retains approximately the same relative importance, regardless of cloze text, when scored by the exact, any-acceptable and grammatically correct scoring procedures. When the cloze is scored by the form class procedures, on the other hand, the increase in importance of the correlation of cloze with the difficult dictation as the text used becomes easier mirrors that of the easy dictation.

On the difficult text, the cloze is more closely related to the Grammar, Vocabulary and Listening Comprehension Tests of ELBA (Tests 5, 6 and 4), the subtotal of Part 2, and the overall ELBA total, than to the dictation tests.

On the medium text, the cloze correlates higher with the part totals and overall total of the ELBA battery than it does with dictation, but it correlates higher with the latter than with any of the individual ELBA tests. The easy text correlates higher with the easy dictation than with any of the ELBA tests, subtotals or overall total.

Table 8.22 removes the Part 1 and Part 2 subtotals, and the overall ELBA total, and simply compares the correlations of cloze: dictation with those of cloze:ELBA individual tests, and in so doing gives a clearer picture of how cloze relates to other separate measures of English-as-a-foreign-language proficiency.

TABLE 8.22

Rank order of correlations of cloze with ELBA individual tests and dictation, by text group.

	ELBA							Dictation	
	1	2	3	4	5	6	7	Easy	Difficult
Difficult Text									
Exact	8	7	9	3	1	2	4=	6	4=
SEMAC	8	7	9	3	1	2	4	6	5
GRCO	8	7	9	5=	1	2	5=	3	4
IDFC	8	4	9	3	1	2	7	5=	5=
ACFC	8	6=	9	3	1	2	6=	4=	4=
Medium Text									
Exact	7	8=	8=	4	1=	5	6	1=	1=
SEMAC	7	8	9	1	4	5	6	2=	2=
GRCO	6	9	8	1=	3=	5	7	3=	1=
IDFC	8	9	7	3	2	5=	5=	4	1
ACFC	7	9	8	3	2	6	5	4	1
Easy Text									
Exact	6=	6=	9	2	3=	5	6=	1	3=
SEMAC	7	8	9	2	3	5	6	1	4
GRCO	7	8	9	4	2=	5	6	1	2=
IDFC	7=	7=	9	3	4	6	5	1	2
ACFC	7	8	9	3	4	6	5	1	2

On the difficult text cloze clearly relates more to "core proficiency" - Tests 4, 5 and 6 - than to dictation. The lowest correlations are with Tests 1, 2 and 3, which have a heavy listening component,

but the fourth listening test - Test 4 - is related to the cloze than dictation is. In fact, the correlations of cloze with dictation are quite low (of the order of .5).

On the medium and easy texts, however, cloze correlates higher with at least one of the dictations than with any of the ELBA tests. This fact is due in part at least to the circumstance that as cloze texts become easier, the cloze correlation with ELBA decreases, whereas the correlation with dictation increases - i.e., the change in text has different effects on ELBA and dictation.

However, there are differences between the two dictations in their correlations with cloze. On the medium text, there is a tendency for the cloze to relate more to the difficult dictation than the easy dictation; on the easy text the situation is reversed. Moreover, even on the easy text cloze does not correlate higher with <u>both</u> dictations than with any ELBA test. Tests 4 and 5 are both closer to cloze than cloze is to the difficult dictation, for two of the scoring procedures, and on the medium text, the same two ELBA tests are closer to cloze than cloze is to the easy dictation, when the cloze is scored by any procedure other than the exact word. It is also noteworthy that the Reading Comprehension Test (ELBA Test 7) almost always correlates lower with the cloze than the dictation does. There is only one occasion - difficult text, any-acceptable scoring procedure - where the cloze correlates higher with Test 7 than with either of the dictations. This confirms the point, frequently made in connection with cloze and dictation, that the cloze correlates better with dictation than with tests of reading comprehension. This does not, of course, necessarily invalidate tests

like ELBA Test 7 as tests of reading comprehension.

Finally, as the previous paragraph has shown, different scoring procedures relate differently to ELBA and dictation, in absolute terms and in relative importance. In particular, as noted, the form class scoring procedures work differently from the other three procedures.

Summary

To summarise these findings, it was found that the easy dictation relates more closely to ELBA and to the cloze than does the difficult dictation, but that the pattern of correlations is approximately the same for both dictations.

The dictations related more closely to an easy text cloze or a medium text cloze than to a difficult text cloze, and more to cloze scored by the any-acceptable-word procedure than by the exact word method. However, dictation was not seen to correlate more highly with cloze than with ELBA, and, in particular, it appeared to be more related to tests of listening comprehension, sound recognition, vocabulary and grammar than to the exact cloze, or to reading comprehension.

Just as in the investigation of cloze and ELBA, this section has shown that several variables affect the relationship of cloze with the dictation. The deletion rate used in creating the cloze test, the text on which the test is based, and the procedure used to score the tests all have a considerable effect on cloze's relationship with dictation. No generalisation is possible about the effect of deletion frequency, other than to say that there is an inconsistent effect. As regards scoring procedures, the any-acceptable-word procedure always

correlates best with the dictation, followed by the GRCO and then the exact word. The influence of the passage on which the cloze is based is shown in that an easy text results in higher correlations with dictation, and in that as the text becomes easier, the cloze relates more to dictation than it does to the ELBA. However, this effect is not true of dictation as a whole, but of the easy dictation. With the difficult dictation there is little change as the cloze text becomes easier, and in fact even the easy cloze correlates higher with some ELBA tests than with the difficult dictation. The difference between the two dictations is also seen in that on the difficult text, the cloze tends to correlate more with the difficult dictation whereas on the easy text it correlates more with the easy dictation.

8.6 Factor analyses of cloze, ELBA and dictation

In an attempt to answer the question "What does the cloze test?" several factor analyses were made of the cloze data together with the other measures of proficiency in English as a foreign language used in the second part of this study, i.e., with non-native speakers of English. So far, no attempt known to the author has been made to factor-analyse cloze as a measure of such proficiency, although it is suggested in the literature, based on correlational studies, that cloze is a measure of English-as-a-foreign-language proficiency. That being the case, one might expect that cloze would measure the same factor of proficiency as existing, presumably valid measures. The present study is purely exploratory, in an attempt to see whether such analyses are suggestive of any hypotheses regarding the nature of the cloze used with non-native speakers, and also to determine which of two possible factorial solutions

is most meaningful. The first solution, or series of solutions, is the traditional factor-analytic method, namely, to perform a principal components analysis with unities in the main diagonal, extract those factors with an eigenvalue equal to or greater than 1.00, and rotate these factors orthogonally, using the Varimax procedure. The second series of solutions is different in that it continues to extract factors in the principal components analysis until all factors with an eigenvalue of at least 0.5 have been extracted. These factors are then rotated in the usual way. Clearly, other methods of factor analysis are possible, particularly the oblique rotation, and subsequent investigators may wish to attempt such analyses.

In both cases, four separate sets of analyses were performed. In the first, each cloze scoring procedure was entered separately into the matrix, along with the ELBA tests. In the second, one dictation - the easy one - was added to the matrix. In the third analysis, the second dictation was included, and finally, all the tests, including all the cloze scores, were entered together into the matrix.

8.6.1 The traditional solution: eigenvalue ≥ 1.0

The results of these analyses are presented in Table 8.23.

If one cloze scoring scheme at a time is entered into the matrix, the results are remarkably consistent for each scoring procedure. On the difficult text, only one factor is present; on the medium text, two factors are identifiable; on the easy text, one factor only emerges when the matrix contains cloze and ELBA, and two emerge when one or two dictations are added to that matrix.

The one factor on the difficult and easy texts loads highly on

all tests. The lowest loading (never below .52) is on Test 3 (Stress), and Tests 1 and 2 (Sound Recognition and Intonation) also tend to load lower than the rest. The highest loading is usually achieved by Test 5 (Grammar). However, these differences are minor in face of the generally high and even loadings for all tests. This factor, which, of course, includes the cloze, can only be called proficiency in English as a foreign language.

The addition of one or both dictations to the matrix for the difficult text has little or no effect on the loadings of each test on this one factor, and dictation itself also loads highly on it. However, it should be noted that the communalities of this unifactorial solution tend to be low, with several tests (usually Tests 2, 3 and sometimes the cloze) going below .6.

The two factors that emerge on the cloze medium text, when dictation is disregarded, seem to be 1) a core proficiency factor, and 2) a suprasegmental, or metalinguistic, or inferential listening factor. The first factor loads consistently for all scoring procedures on Tests 1, 4, 5, 6 and cloze, and somewhat lower on Test 7 (Reading Comprehension). The second factor loads mainly on Tests 2 and 3, with a much lower loading on certain other tests, mainly 4 and 7, and, sometimes, cloze. This solution is more complete than the difficult text and easy text solutions, with no communality going below .6, except that of the cloze score on the acceptable-form-class procedure. Although cloze is more strongly associated with the first factor (loadings range from .67 to .81), a considerable loading is also present on the second factor (ranging from .31 to .41). This complicates the identification of this

second factor, and suggests that the solution is not entirely satisfactory.

The addition of one or two dictations, as noted above, has no effect on the difficult text solution, which remains unifactorial. On the remaining two texts, two factors emerge, but these are even less clearly separate than the previous two factors without dictation. As was noted initially, the cloze scoring procedure used in the analysis does not change the nature or number of factors that emerge in any way. Basically, the two factors seem to be a "with dictation" and a "without dictation" factor; on the "with dictation" factor, reasonably high loadings are present for Tests 1, 4, 5 and 6, and the cloze. The difference between the two text groups is merely that Vocabulary (Test 6) seems to be more important than the other tests on the easy text, whereas on the medium text, Sound Recognition (Test 1) is somewhat more important. Also, the loadings of the cloze are higher on the easy text than on the medium text.

The second factor, "without dictation", is similar to the second factor mentioned in the previous result (medium text) in that the highest loadings are on Tests 2 and 3, with lower loadings on Tests 7, 4 and 5. In fact, each test in the matrix has at least a moderate loading on this factor, except for the dictations. Hence its name. It is interesting to note that, apart from Tests 3 and 2, which load almost exclusively on this factor, Test 7 (Reading Comprehension) tends to load almost as much on Factor 2 as on Factor 1, and on occasions more so.

Again, as far as the cloze is concerned, this solution is not entirely satisfactory. Although the loadings for cloze are always higher on Factor One than on Factor Two, they are never outstandingly high, and

are usually lower than the loadings of at least two ELBA tests. (On the medium text, the loadings range from .53 to .75; on the easy text from .68 to .79.) A substantial amount of variance on the cloze is unaccounted for by this first factor, and especially on the medium text, cloze loads considerably on the second factor also (medium text from .41 to .56; easy text lower at .16 to .39).

Whether one or two dictations are added to the initial correlation matrix makes very little difference to the nature of the factors that emerge. As far as the cloze is concerned, the loading on Factor One is somewhat reduced when two dictations are present, and that on Factor Two is slightly increased.

In summary of this section then, it appears that for the medium and easy texts, two factors are present when cloze, ELBA and dictation are considered. Both factors can be considered proficiency factors, since the ELBA tests load on both, with a tendency for Factor One to resemble the core proficiency factor previously identified, and Factor Two to resemble the second factor - suprasegmental, metalinguistic, inferential listening. The most characteristic thing about the two factors is the presence and absence of dictation. The dictation factor is seen as being associated with core proficiency and cloze, but much variance of these latter elements is also present in Factor Two. No satisfactory explanation of "what cloze tests" has yet been achieved, but it is beginning to look as if cloze is factorially complex.

The factors outlined so far are as follows:

Cloze and ELBA

Difficult Text	Medium Text	Easy Text
Factor 1: all tests, lower on 2 and 3	Factor 1: 1, 4, 5, 6 & cloze	Factor 1: all tests
	Factor 2: 2 & 3; some 4, 7 & cloze	

Cloze, ELBA and Dictation

Difficult Text

Factor 1: all tests, lower on 1, 2, 3

Medium Text

Factor 1: Dictation, 1, 6, 5, 4 and cloze

Factor 2: Tests 2 & 3 & 7, with some 4, 5 and cloze

Easy Text

Factor 1: Dictation, 1, 6, 4, 5 and cloze

Factor 2: Tests 3, 2 with some 5, 4 and 7, little cloze

In an attempt to extract a cloze factor, all five cloze scores were entered together into the matrix, factors with an eigenvalue equal to or greater than 1.00 extracted, and these factors rotated, as before. For the difficult text, this results in two factors, instead of one, and for the other two texts, in three factors instead of two.

On the difficult text, the first factor is a cloze factor, loading highest on the IDFC, ACFC and GRCO scoring procedures. On this factor, there is some loading from Test 5 (Grammar: .61), Test 6 (Vocabulary: .51), and dictation (.45/.38). The second factor seems to be some sort of listening factor, loading on Tests 1, 4 and the difficult

dictation. The cloze is also present on this factor, but at a low level (.3).

On the medium text, the first factor is a dictation factor, with Sound Recognition (Test 1) also present, and some 4, 5 and 6 (Listening, Grammar, and Vocabulary). The second factor is a cloze factor, with only slight loadings for Tests 6 and 5, and the third factor is the previous mysterious second factor, loading on Tests 2, 3 and 7. It should be noted that cloze also loads, although only slightly, on the two non-cloze factors.

On the easy text, the third factor remains the same (Tests 2 and 3, with some loading on Tests 4, 5 and 7), but the order of the first two factors has changed, such that the second factor is the dictation cum ELBA factor (loading on Tests 6, 1, 4, 5 and 7) whilst the first factor is the cloze factor, with some dictation. Again, the cloze loads on the two non-cloze factors.

Summary (all tests)

Difficult Text

Factor 1: cloze, with some 5 and 6, and slight dictation

Factor 2: Tests 1, 4 and dictation, with low 5, 2, 6 and 7

Medium Text

Factor 1: dictation, Test 1

Factor 2: cloze

Factor 3: Tests 2, 3 and 7

Easy Text

Factor 1: cloze, some dictation

Factor 2: dictation, 6, 1, 4, 5, 7

Factor 3: Tests 2 and 3, with some 4, 5 and 7

It has thus proved possible to isolate a cloze factor, which is not particularly associated with other factors, but sometimes weakly associated with dictation and core proficiency.

8.6.2 A second solution: eigenvalue $\geq$ 0.5

All these factor analyses suffer from the defect of low communalities, in other words, much variance in most of the variables is left unexplained. To attempt a more complete solution, it was necessary to continue extracting factors with eigenvalues below the normal end point of 1.00. In this part of the investigation, factor extraction continued with all factors until an eigenvalue of .5 had been reached. This resulted in greater communalities for all the variables. It is difficult to know which solution to choose, and the final decision is inevitably judgmental, balancing the advantages of achieving higher communalities on the one hand against the disadvantage of having factors of less generality and importance. As this section of the study is frankly exploratory, it is perhaps unnecessary at this stage to decide for one or the other solution per se, but more advisable to see which solution results in a more intuitively or theoretically satisfying or suggestive factor structure. The results of the analysis are presented in Table 8.24.

The relatively simple picture that prevailed under the previous conditions (EV $\geq$ 1.0) are somewhat more complicated here. It should be

noted that, whereas for both difficult and medium texts the number of factors that are extracted is the same for all scoring procedures, on the easy text the SEMAC and GRCO procedures result in one factor fewer than the other three scoring procedures. Further, whereas on the medium and easy texts the addition of two rather than one dictation to the matrix has no effect on the number of factors that occur, on the difficult text one more factor emerges, for four of the five scoring procedures, when the second (difficult) dictation is included.

There are, then, the following numbers of factors present:

Medium Text - with and without dictation: four factors

Difficult Text - without dictation, and with one dictation: four factors

- with two dictations: five factors

Easy Text - for the exact, IDFC, and ACFC procedures, three factors

for the SEMAC scoring procedure - with and without dictation: two factors

for the GRCO scoring procedure - without dictation: two factors

for the GRCO scoring procedure - with dictation: three factors

More important than the number of factors, however, is their nature.

To begin with the simplest case first, the medium text, the cloze scoring procedures largely agree with one another as to the factor structure of the various variables. The only exceptions are the form class scoring procedures, which tend to change the order of importance of some of the factors. As their components are substantially similar

to the factors for the other scoring procedures, the differences are ignored here.

When cloze and ELBA only are examined, the factors are as follows:

Factor 1: Sound recognition (Test 1) and cloze, with some core proficiency

Factor 2: Intonation or inference (Test 2)

Factor 3: Reading and listening comprehension (Tests 7 and 4), with some core proficiency

Factor 4: Sentence stress (Test 3)

When dictation is added to the matrix (since the presence of only one dictation makes little difference, this dimension will be ignored), the factors do not substantially change, except for Factor 1, which is now composed largely of dictation and sound recognition, and the importance of cloze is reduced (from a range of .71 to .78 down to .40 to .65). Cloze now tends to load somewhat higher on the second factor, which remains, however, basically an intonation/inference factor. Factors 3 and 4 are unchanged. As was seen previously, cloze is not solely represented on one factor, but is spread over at least three of the four factors. It is most interesting to note that it does not load highly on the comprehension factor (Factor 3), and that when dictation is added, it tends to remove the cloze element from Factor 1. The relationship between cloze and dictation, and cloze and comprehension, appears to be less close than previous (correlational) studies suggest. It is difficult, however, on theoretical grounds, to account for the close association shown between cloze and sound recognition, and cloze and Test 2 (intonation/inference).

When dictation is added to the matrix, then, the factors emerge as follows:

Factor 1: 1 , dictation, some cloze, 6, 5, and 4

Factor 2: Intonation (2) and cloze, some 6 and 5 (vocabulary and grammar)

Factor 3: Reading and listening comprehension (7 and 4)

Factor 4: Stress (3)

Although on the easy text, as on the medium text, there is little effect of including one or two dictations in the analysis other than some change in the order of importance of two factors, there is some disagreement among scoring procedures as to the number and therefore the nature of factors involved. When three factors emerge from the analysis on ELBA and cloze alone, the structure of the first two factors is similar to that of the medium text ($EV \geq 1.0$); in other words, Factor 1 is a general proficiency factor, loading on Tests 6, 1, 4, 5 and 7, and Factor 2 is a listening factor, loading on Tests 3 and 2. The difference is that cloze does not load on Factor 1, but instead on its own factor, Factor 3, which is exclusive to cloze. However, in those cases (SEMAC and GRCO) where only two factors are present, this has the effect of causing cloze to load moderately high on the first, general proficiency factor. These facts are suggestive of a separate cloze factor whose existence is hidden by its association with a general proficiency factor if the analysis does not go far enough.

When dictation is added to the matrix, what happens is that when, as on the SEMAC, only two factors are present, the dictation loads high on the first, general, core proficiency factor. When there are

three factors, however, it tends not to load on this first factor, but to be much more closely associated with the cloze factor.

To summarise so far, on the easy text there are at least two factors, a proficiency factor (with dictation and cloze loading high only when only two factors are extracted) and a listening factor. A third factor, usually present, except for the SEMAC, is the cloze factor, on which only the dictation, when present, loads. These results contradict the results of the medium text in that

a) there is no comprehension factor

b) unlike the medium text, a general proficiency factor is present

c) a separate cloze factor exists, which associates with dictation to reveal a greater degree of relationship than shown by the medium text.

It is, of course, possible, in part at least, that the different results are due to the number of factors which emerge from the different texts. As the difficult text sometimes shows five factors, it would be expected that, if the number of factors has an effect, the analysis will be different again.

On the difficult text, when cloze and ELBA alone are analysed, the scoring procedures agree both on the number of factors and their structure. Four factors are present, and cloze loads mainly on the first one, which is the core proficiency factor. Tests 6, 5 and 4 as well as the cloze are all closely associated with this factor. The second factor seems to be a (mainly) segmental, discrete listening factor with loads on Tests 1 and 2. The third factor involves Tests 7 and 2, with some cloze - the cloze loading diminishes as the scoring procedures become less

demanding. As Test 2 (Intonation) demands a good reading ability, it seems not unreasonable to label this factor a reading comprehension (inferential) factor. The fourth factor is another listening factor, on Test 3 alone.

To summarise these factors:

Factor 1: Core proficiency (6, 5, 4) and cloze

Factor 2: Discrete listening (1 and some 2)

Factor 3: Reading comprehension (inferential) (7 and 2)

Factor 4: Sentence stress (3)

When one dictation is added to the matrix, the factor structure does not change, except for the cloze exact scoring procedure. With that exception, the dictation loads on the core proficiency factor along with the cloze, and, to some extent, on the segmental listening factor (now Factor 3). However, when cloze is scored by the exact word procedure, five factors appear - the extra factor being, interestingly, a cloze/dictation factor. The four other factors remain unchanged, apart from the absence of loadings on cloze and dictation.

When both dictations are added, the fifth factor is maintained, but with an important difference. This is that the cloze/dictation factor becomes a dictation factor, with some cloze only. Most of the cloze loading moves back to the core proficiency factor, particularly on the form class scoring procedure. Unlike the dictation, which loads almost exclusively on its own factor, the cloze loads at least moderately on several factors, none of which is identifiably a cloze factor. The five factors that emerge are as follows (the order of the factors varies from scoring procedure to scoring procedure but their contribution remains

the same):

Factor 1: Core proficiency (mainly Tests 6, 5, 4, and cloze)

Factor 2: Dictation (with some cloze)

Factor 3: Reading comprehension (mainly Tests 7 and 2, sometimes cloze)

Factor 4: Segmental listening (Tests 1 and 2)

Factor 5: Stress (Test 3)

The prediction that the nature of the factors would change as more factors emerge has been borne out on the difficult text, which has resulted in factors which are dissimilar to the factors of the easy text, and, to some extent, to those of the medium text.

As the cloze becomes more difficult, it appears that the monolithic proficiency factor is broken up, with the dictation, the segmental listening test, the reading comprehension test, and the stress test separating out into their own factors. Unlike the dictation, the cloze clearly remains associated with this core proficiency factor. The cloze/dictation factor seen on the easy text has turned into a dictation factor on the difficult text, with some cloze association only. The evidence from the difficult text points to a weak association of cloze with both dictation and reading comprehension, and a major association with core proficiency (grammar, vocabulary and general listening).

Finally, as on the previous analysis, an attempt was made to extract a cloze factor by entering all five cloze scores into the matrix, extracting and totalling all factors with an eigenvalue of at least 0.5. For the difficult text, this results in six factors; for the medium text

five factors; but the easy text remains unchanged at three factors, since no factor had an eigenvalue of between 1.00 and 0.5.

It will be remembered that the three factors on the easy text were:

Factor 1: Cloze, with some dictation

Factor 2: Dictation and ELBA, except Tests 2 and 3

Factor 3: Tests 2 and 3, with some grammar and comprehension

When more factors are extracted, on different texts, the situation changes, and, in particular, dictation loads on its own factor. On the medium text, the first factor is a dictation factor, with loading also on Test 1, segmental listening. Cloze has very little association with this factor, and, in fact, loads on its own factor, the second, with virtually nothing else. The overall factor structure is as follows (it should be noted that cloze only loads, and that only slightly, on Factor 1, apart from Factor 2. The exact word score is the only scoring procedure which has any loading on Factor 3 - suggesting, indeed, that it is somewhat different from the other procedures):

Factor 1: Dictation and Test 1

Factor 2: Cloze

Factor 3: Test 2

Factor 4: Tests 7 and 4 - comprehension

Factor 5: Test 3

The difficult text tends to produce more separate factors. The dictation, in particular, is now isolated on its own factor, unrelated to other variables. Again, the cloze factor appears, though with some association with grammar. The factors are as follows:

Factor 1: Cloze, some grammar

Factor 2: Dictation

Factor 3: Tests 1 and 2

Factor 4: Test 7

Factor 5: Test 6

Factor 6: Test 3

The exact and SEMAC cloze scores have a slight loading on Factors 2 and 4. The other procedures load only on Factor 1.

8.6.3 Summary and conclusions

The difference between the two factor analysis solutions (eigenvalue greater than 1.00 and 0.5 respectively) can be summarised as follows:

1) On the 0.5 solution, more factors emerge, and the communality is greater.
2) The factors are consequently of less generality on the 0.5 solution.
3) On the 1.00 solution, cloze is shown to be associated with both the dictation and core proficiency in English as a foreign language, but the solutions are far from complete.
4) The 0.5 solution is less monolithic or unitary, and therefore more explanatory and intuitively satisfying. Cloze is shown to be factorially complex, as only on the easy text does a cloze factor emerge when only one scoring procedure at a time is considered. There is a suggestion that as more factors are used, it is more likely that a specific cloze factor will emerge, at least on the easy text.

The results of the analyses as far as dictation is concerned are as follows:

1) Dictation is associated with some of the ELBA tests at least as often as with the cloze.
2) On one analysis (EV $\geq$ 1.00), dictation is associated with both proficiency in English-as-a-foreign-language, and with cloze.
3) On the second analysis (EV $\geq$ 0.5) the dictation either loads on its own factor exclusively, or with either the cloze or sound recognition, and not with proficiency in English as a foreign language.

The summarised points which follow regarding the cloze should be read in the light of several qualifications:

1) Different texts used for the cloze test result in different factors. On the second analysis, in particular, on an easy text, a pure cloze factor was present. On the medium text, cloze and intonation loaded on the same factor, whereas on the difficult text the cloze loaded on several factors, but particularly on core proficiency - vocabulary, grammar and general listening comprehension.
2) Although different scoring procedures do not result in different factors, the amount of their loading varies. In particular, the any-acceptable-word procedure loads highest on relevant factors. Some suggestion was found that as the scoring procedures become more demanding - i.e., from "any acceptable form class" to "exact" - the cloze becomes more of a reading test.

The results of the factor analyses show cloze to be factorially complex. A factor exclusive to the cloze was occasionally present, especially as more factors were entered into the analysis. Furthermore,

as one might expect, when all the cloze scores are included in the initial matrix, a cloze factor on which little else loads clearly emerges.

More frequently, the cloze tended to be associated with dictation on the same factor. However, this association with dictation by no means completely explained the cloze, as it not only loaded elsewhere, it frequently was not associated with the dictation to any significant extent. Most frequently, the cloze loaded on a factor which is best labelled "core proficiency", composed of ELBA Tests 4, 5 and 6. Perhaps the most important finding is that cloze is less associated with reading comprehension, and especially tests of inferential reading ability, but closely associated with tests of grammar and vocabulary.

In answer to the initial question as to whether cloze measures the same "proficiency in English as a foreign language" factor as the ELBA, the evidence from this study is that, on the whole, the cloze <u>is</u> such a measure. The unifactorial solution shows cloze to load high on a general proficiency factor. When two factors are extracted, the cloze is associated with core proficiency, and even when five factors are extracted, the cloze is still seen to associate with this factor. In addition, the study confirms some studies and refutes others to show that the any-acceptable-word scoring procedure is the best measure of the proficiency factor, but, further, that almost any scoring procedure is capable of measuring this factor.

Nevertheless, it remains true that cloze is not confined to this proficiency factor, and the nature of the other factors with which it associates is more obscure.

CHAPTER 9

Summary and Discussion

9.1 Non-native speakers

9.1.1 Algeria - summary

The pilot study in Algeria posed two questions for investigation: 1) Is the cloze procedure unitary - in other words, is it possible to generalise confidently about the nature of the cloze as a technique for producing tests and exercises of reading comprehension and language proficiency? 2) (the same question from a different angle) Are different versions of the cloze test comparable? That is, do different cloze tests measure comprehension, readability or linguistic proficiency differently?

Two main variables were investigated. The effect of varying the frequency of word deletion from text was examined to see if and how this would affect the measurement of readability and estimates of English proficiency. Varying the difficulty level of text was also investigated for a possible effect on the estimate of proficiency. Furthermore, the relationship between two measures of readability for students of English as a foreign language - cloze and the Fog formula - was also examined.

Significant differences among deletion frequencies were found, often in the predicted direction - namely, that more frequent deletion would result in a more difficult test and, therefore, a different measure of the readability of the text. It was found that deletion rate four (every fourth word deleted) was always more difficult than the rest,

but that on some analyses deletion rate six was also significantly easier than deletion rate four and significantly more difficult than the less frequent deletion rates. A tendency was also discernible for deletion rate 8 to be more difficult than deletion rates 10, 12 and 14. It was suggested, therefore, that perhaps at least nine words of context are necessary on either side of a blank before one can regard the effect of varying deletion frequency to have been eliminated.

It was expected that good students would be less affected by variation in amount of context than would others, and this proved to be the case. However, the converse expectation - that weak students would need more context in order to complete the blanks - was not confirmed. In other words, an increase in context does not help even poor students to predict missing words.

However, these conclusions were based on results which grouped the different texts together, and assumed the effect of their difference to be negligible. When text differences were taken into account, there was virtually no agreement on the difference between deletion rates, since on some texts all deletion rates gave the same results, on others different results were gained from each rate, and on still others only one or two deletion rates were different from the others. It was impossible to predict the effect of varying deletion frequency for any one text.

Although the more frequent deletion rates tended to rank texts in the same order of difficulty, the less frequent deletion rates differed in their estimates of relative difficulty. If this difference is real one, then an interaction between text and deletion rate as measures of

readability is indicated. In so far as there is an interaction, then the deletion rate used is an important variable in readability studies. Of course, if there is no interaction effect then different results would not be expected when different texts are used to investigate the nature of contextual constraint or the effect of varying deletion frequency. The Algerian study, however, does not provide a clear answer to this point.

The cloze technique did not prove capable of distinguishing any two texts. In the Algerian study, no one text emerged as being consistently significantly different from any other, and when cloze scores descended to around 20% correct restorations, the differences between texts were minimal.

As regards the relationship between cloze and the Fog readability formula, it was found that there was no correlation whatsoever between the two procedures in measuring the readability of text. If Fog is a valid measure of readability for native speakers and cloze an equally valid measure for non-native speakers, it would appear that the nature of the difficulty of text is different for the two populations. It was suggested that a measure of sentence length is not an adequate measure of sentence complexity (or difficulty) for adult learners of English as a foreign language who are already fluent readers in their own language. It was also indicated that perhaps the non-native speaker's familiarity with a wide vocabulary in his native tongue might mean that the difficulties presented in text by lexis are qualitatively different from the difficulties native speakers encounter and which are supposed to be measured by a count of the syllables comprising the words.

The effect of varying the two main variables - text and deletion rate - on the measurement of English language proficiency was also investigated. No clear-cut results emerged. No consistently higher correlations for any text with the criterion proficiency measure were observed. There was, however, a slight tendency for cloze tests based on difficult texts to correlate higher with the criterion that did easy cloze tests. Variations were observed in the correlation of different deletion rates with the criterion, but no one deletion rate emerged as a better predictor or measure of English proficiency than any other. It was suggested that more frequent deletion should tax students' syntactic abilities more than less frequent deletion, but no evidence was found that might have supported this, since the less frequent deletion rates did not correlate less with the subsections testing grammar than the more frequent deletion rates. The evidence indicates that whatever cloze tests, it is measured more or less equally by any deletion rate. However, this conclusion was put in doubt by the low intercorrelation obtained between deletion rates. If they were testing the same thing, one would expect high intercorrelations. That this was not the case suggests there might indeed be an interaction between deletion rate and text difficulty in the use of cloze to measure English proficiency. This supposition is supported by the widely varying correlations with the criterion gained by individual cloze tests.

Finally, it was found that the overall relationship between cloze and proficiency in English as a foreign language was only moderate, and much less than previous studies had indicated. Moreover, the highest correlations were not achieved with supposedly integrative tests, but

with traditional discrete-point grammar tests. This pattern was true regardless of the text or deletion rate variables.

It was felt that further research was needed into the effect of changing the variables, and that the as-yet-uninvestigated variable of scoring procedure should also be examined. A different measure of English proficiency, explicitly covering more aspects of proficiency (for example, vocabulary and reading comprehension) as well as a reputedly integrative test technique - the dictation, said to relate closely to the cloze, to language proficiency and to reading comprehension - were used to facilitate in-depth examination of the cloze procedure in the main study.

9.1.2 The main study - summary

9.1.2.1 Experimental variables

The Algerian study having shown that the cloze could not discriminate among relatively similar texts for non-native speakers, the main study investigated whether the procedure was capable of distinguishing obviously different texts from one another. The clear result of the investigation was that the cloze procedure always ranked the three texts (intended to be easy, medium and difficult) in the same order, which was the order assumed in advance to be the correct order of difficulty, regardless of the experimental variables. No matter which deletion rate was used, or which scoring procedure (even the least demanding procedures), the texts were consistently ranked in the same order. For gross discrimination of readability, the cloze is apparently adequate, although its ability to make finer distinctions is still in doubt.

There was a significant interaction between the text variable

and the deletion frequency, so that what was true for a deletion rate on one text was not true for the same deletion rate on another text, and different patterns of relationships among deletion rates were seen for different texts. The Algerian study had found deletion rate 4 always to be more difficult than the other deletion rates, so this frequency was not included in the main study. Deletion rate 14 had virtually always been the same as deletion rates 10 and 12, and so no further investigation of this frequency was made in the main study. The frequencies studied were, therefore, 6, 8, 10 and 12.

Significant differences were obtained between certain deletion rates on the difficult and easy texts. However, on the medium text no differences between deletion frequencies were gained. Moreover, the existence of significant differences varied from scoring procedure to procedure. When significant differences between deletion rates were obtained, they were not consistent. For instance, deletion rate 6 was found to be significantly different from deletion rate 12 on one text, but not on another. In one case, deletion rate 8 was different from deletion rate 6; in another case there was no difference. The only consistent finding was that deletion rate 10 was never different from deletion rate 8. Similarly, there was no consistency in the direction of the differences that were found. On the difficult text, deletion rates 8 and 10 were more difficult than 6 and 12, whereas on the easy text deletion rate 12 was, counter-intuitively, more difficult than the other three deletion frequencies.

In summary, text differences cannot be ignored, since there is a significant interaction between deletion rate and text. This means

that the results of the Algerian study, showing increasing ease as the deletion frequency decreased, must be questioned, since they were obtained by ignoring text differences. When text differences are taken into account, it is impossible to conclude that more frequent deletion results in more difficult tests. The differences between tests, therefore, are almost certainly not due to the different quantity of context surrounding each blank, or, to put it another way, increased context in a cloze test does not seem to result in greater predictability of missing words.

A further investigation suggested a possible solution to the problem. When only identical items, deleted from both cloze tests under consideration at any one time, were compared, no significant differences whatsoever were found. This provides confirmation that varying the amount of context round a blank has no effect on that blank, and indicates that the reason for the difference between individual cloze tests is simply that different deletion frequencies delete different words, which may be easier or more difficult, according to chance.

Examination of the scores produced by the five scoring procedures showed that they were almost always significantly different from each other. In particular, those procedures designed to permit more constraint to operate on each blank resulted, as expected, in significantly lower scores. Nevertheless, there was a high degree of agreement among procedures on the ranking of subjects, since over one third of the of the intercorrelations were above .9, and the lowest, out of a total of 120 coefficients, was .71.

Of these interrelations, the closest were among the three

grammatical scoring procedures, with almost perfect coefficients being attained. The conclusion from this would be that whatever the different grammatical procedures are intended to test, in practice it makes no difference whether one scores simply for correct form class replacement, or whether one takes into account other grammatical features like tense, concord, number, etc.

The any-acceptable-word procedure correlates as closely with the grammatical procedures as with the exact word procedure, but the exact word procedure correlates higher with the any-acceptable procedure than with any other procedure. This latter finding might be taken as indicating that the ability to fill in a blank with a grammatically appropriate word - however "appropriate" may be defined - is relatively unrelated to the ability to replace the exact word, which might be thought to demand sensitivity to style, author's intention and the like. However, virtually all of the coefficients were of the order of .80 or above, which suggests that such a conclusion would be creating a false dichotomy. In fact, the different scoring procedures - even those superficially most *different - are very closely related to each other, and little* evidence has been found to suggest that different procedures measure different abilities. This question will be examined more closely when cloze is compared with other measures of proficiency in English as a foreign language. It is debatable, however, whether the coefficients are high enough for one to use one procedure in place of another. Even the correlations between the exact and any-acceptable procedures do not necessarily justify the replacement of one by the other, although, as pointed out in Chapter 7, they do confirm previous research findings that

there is a close relationship between the two.

Finally, it has already been suggested that there is an interaction between scoring procedures, deletion rates and text. Different combinations of text and deletion rate will result in different intercorrelations of scoring procedures. It was found that different scoring procedures have different effects on the experimental effect of other variables. In particular, the any-acceptable procedure resulted in significant differences between deletion rates on only one text, whereas the exact word procedure showed significant differences between deletion rates on two texts. In other words, to reduce the effect of varying the deletion rate, it would be sensible to score for any acceptable word, and not just the exact word.

The efficiency of the cloze as a test was looked at from the point of view of traditional test analysis. Considerable doubt was thrown on the tools of such analysis, and their applicability to the case of the cloze procedure. No satisfactory measure of reliability was found which could allow for the great variation in items difficulty typical of the cloze test. Indeed, it is doubtful to what extent cloze blanks can be considered test items and analysed with the usual item statistics in a way which suggests that one item is essentially separate from, and independent of, any other.

It is difficult to see, moreover, how to improve a cloze test, basing one's supposed improvement on traditional analyses, if the revision itself is not going to change drastically the nature of the cloze. In short, it was not at all clear how valid it is to analyse one testing procedure using instruments designed essentially for different procedures.

Nevertheless, in order to compare the cloze with other procedures, it is necessary to have a common yardstick and so, regardless of the deficiencies, the cloze was examined traditionally.

Although no one text appeared to give better reliability than another, there was a clear interaction between deletion rate and text on reliability, as on many other variables, so that different deletion rates gave unpredictably different reliabilities.

Both the difficult and medium texts, when scored by the exact word method, gave very narrow score distributions. These distributions were dramatically improved when the tests were scored for any acceptable word. In fact, the any-acceptable-word procedure resulted in better distributions than the other scoring procedures, better reliabilities, and less frequent attainment of extreme scores.

Although the exact word method resulted in the best distribution of item difficulty on the easy text, on the difficult text it resulted in the worst distribution. Again, the any-acceptable-word procedure gave the best item facility distributions for both difficult and medium texts. However, it was noted that regardless of scoring procedures, many inefficient items resulted from any cloze test - a maximum of 60% of all items proved to be satisfactorily easy or difficult. Similarly, cloze was relatively poor at producing items capable of discriminating among subjects, although the any-acceptable procedure achieved fewer negative discriminations and more positive discriminations than the exact word procedure.

The general result of the analysis seemed to be that, in traditional testing terms, the cloze is relatively inefficient and not

susceptible of manipulation for improvement. It was recommended that if the exact word scoring procedure was to be used, then an easy text should be selected as the basis for the test, whereas if it was desired to use a relatively difficult text for the test, then the subjects' responses should be scored by the any-acceptable-word procedure.

9.1.2.2 Cloze and English-as-a-foreign-language proficiency

Since all cloze test groups were equivalent for proficiency in English as a foreign language, as traditionally measured, it was assumed that if the cloze tests proved to be different one from another, it would be because the cloze measured something else, or the same thing in a different manner. The results showed that the medium text never revealed differences among groups, nor did the any-acceptable-form-class scoring procedure. However, on the easy text there were differences for two scoring procedures, including the exact word, and on the difficult text significant differences among groups were found when the tests were scored by four of the five scoring procedures. These findings lead to the suspicion that different versions of the cloze test measured proficiency differently, or measured it less successfully.

With regard to the experimental variables, the following findings were made.

9.1.2.2.1 Deletion rate

Different deletion rates relate differently to the ELBA; however, no consistent pattern emerged. (For example, deleting every sixth word resulted in the lowest coefficient on one text, and the highest on another.) Different deletion rates relate differently to the ELBA subtests. Test M8 correlated highest with Sound Recognition, Test E12 with

with Intonation, whilst Test E6 correlated highest with Reading Comprehension. Each test appeared to measure something different from the next, making it extremely difficult and misleading to generalise about "what cloze tests". Moreover, because of the lack of patterning across texts for different deletion rates, it is not possible to conclude that a more frequent (or less frequent) deletion rate gives a better, or different, measure of proficiency in English as a foreign language than another deletion rate.

9.1.2.2.2 Text

Different correlations were achieved with different texts. In general, the highest correlations with ELBA were achieved with the difficult text, and the lowest with the easy text. In particular, the difficult text proved the best predictor of ELBA Tests 5 and 6 (Grammar and Vocabulary). Although each text seemed to present more or less the same profile of foreign language skills, there was a tendency for grammar and vocabulary abilities to be less associated with the cloze as the text became easier. It was suggested that a difficult text might be seen as measuring more the elements of linguistic proficiency, whereas easier texts relate more to global skills like comprehension, although, notably, not the reading comprehension measured by ELBA Test 7.

9.1.2.2.3 Scoring procedures

A clear interaction was observed between scoring procedures and deletion rates, and scoring procedures and text. The use of the exact word procedure tended to emphasize the differences among deletion rates and texts as measures of English-as-a-foreign-language proficiency, whereas the use of the any-acceptable-word method or the grammatical

procedures reduced these differences considerably. The closest relationship with ELBA proficiency, in part or in total, was gained by the any-acceptable procedure, whilst the exact word procedure resulted third best. Nevertheless, no evidence was found to suggest that different scoring procedures actually measured different aspects of proficiency. In particular, it was not the case that grammatical scoring procedures related more to tests of grammatical skills than did other procedures.

In general, the cloze related more closely to the tests of core proficiency in English as a foreign language than to tests of reading comprehension.

9.1.2.2.4 Cloze and dictation

Because of frequent claims that cloze relates more to an integrative test of global skills like dictation than to traditional discrete-point tests, the relationship of cloze and dictation was examined and compared with that of cloze and ELBA. Cloze was not found to relate consistently more to dictation than to ELBA. In fact, the dictation related more to listening and sound recognition tests, and tests of vocabulary, than to the cloze. As the cloze text became easier, there was a tendency for the cloze to relate more to one dictation (the easy one) than to ELBA; however, even on the easy text, the cloze related more to some ELBA subtest than to both dictations. Different scoring procedures and texts resulted in different relationships with the dictation, just as they had with the traditional proficiency test. As on the ELBA, the any-acceptable-word and the grammatically-correct-word procedures were better predictors of dictation than the exact word procedure. However, the level of correlations of cloze with dictation varied greatly (for

example, the any-acceptable-word procedure varied from .38 to .91).

Thus, previous findings were not confirmed that the cloze relates more to integrative tests than to discrete-point tests, or to tests of overall skills rather than tests of linguistic elements. Contrary to some previous research, it was found that the dictation itself related more to tests of grammar and vocabulary than to tests of comprehension, regardless of the integrative/discrete-point dichotomy. It was suggested that in fact the dictation measures lower-order skills rather than the higher-order skills of inference and the like measured in the comprehension tests. Thus the degree of integration of a test was felt to be irrelevant to an analysis of what the dictation measures.

9.1.2.2.5 Factor analysis

Finally, in an attempt to discover indications of what the cloze measures, a series of factor analyses were carried out. Different texts were found to result in different factors, but different scoring procedures did not - they simply resulted in different loadings on the same factors. Cloze was seen to be associated both with dictation and with core proficiency, yet it was felt that the solutions discovered were incomplete. For this reason, it was suggested that cloze is factorially complex. Only on occasions did a separate cloze factor emerge, although it is probable that as more tests are included in the analysis, a specific cloze factor is more likely to emerge. In particular, it is likely that a cloze factor is hidden behind the proficiency factor because of the lack of delicacy of the analysis.

Cloze was most frequently associated with core proficiency and rarely with reading comprehension - especially with inferential reading

abilities. The tests that were thought to measure these abilities loaded on their own factor. Although on a coarse analysis the dictation associated with proficiency and the cloze, a specific dictation factor tended to emerge as the analysis became finer.

In summary, the analysis gave a unifactorial solution to the cloze, showing it to be closely related to core proficiency, but there was a suggestion that the real situation was hidden by the analysis, that cloze is somewhat more complex, and that a complete solution has yet to be found.

9.1.3 Discussion

9.1.3.1 Text

Although the Algerian study showed that the cloze procedure does not reliably distinguish relatively similar texts, the cloze tests of the main study constantly differentiated between obviously different texts. These findings confirm results from studies like Haskell (1973) and Taylor (1953) that cloze is capable of discriminating texts regardless of the deletion rate or the scoring procedures used. When texts are obviously different the cloze will distinguish them. However, when the differences are not so great, consistent discrimination will not be achieved. (This fact reflects the findings of Mosberg et al (1968) that cloze could not satisfactorily distinguish passages at similar low levels of difficulty.) It would seem important to bear text differences in mind when carrying out cloze research. In particular, since a clear interaction was found between deletion rate and text, it is important to take the text variable into account when carrying out research on the effect of contextual constraint by means of cloze tests. Moreover, it

is misleading to ignore differences in text by grouping scores across different texts together in order to examine the effect of some other variable. It is particularly important to bear this in mind when interpreting, for example, the Algerian study's or MacGinitie's (1960) results of the investigation of deletion rate differences (which conclusions were gained by ignoring text differences).

Both the Algerian and the main study discovered that, although on the whole there was little difference between texts as predictors of proficiency in English as a foreign language, there was a tendency for the difficult text to be more related to such proficiency. This confirms Carroll et al's finding that different texts result in different correlations with proficiency in a foreign language. Darnell (1968) also found that non-engineering texts tended to relate more to proficiency than did engineering texts for students of engineering. To the extent that engineering students are likely to find non-engineering texts harder than engineering texts, this represents the same result.

Similarly, Oller (1972) found lower correlations for a very easy text with his criterion than for more difficult texts.

This study has shown a tendency for the difficult text to be more closely related to tests of grammar and vocabulary than are the easier cloze texts. Moreover, the relationship between cloze and dictation varied according to the text used. When an easy text was used, the cloze related more to the easy dictation, whereas with a difficult text, the cloze related more to the difficult dictation. An explanation for these findings is somewhat problematic. It could be that a difficult cloze is a better measure of linguistic elements (syntax and lexis)

because it demands a greater command of such basic linguistic skills. In other words, an easy text can be clozed by students whose command of such basic skills is faulty, but adequate to cope with uncomplicated text. An easy cloze would thus be a less efficient test of such skills. The same might be said for a difficult dictation - that the student needs a good command of vocabulary and grammar in order to be able to handle it. However, the results of the correlation of the dictations with the ELBA indicate that the easy dictation is a better predictor of English-as-a-foreign-language proficiency. This result is somewhat confusing, but there is still some attraction in the explanation that a difficult cloze text is more likely than an easy text to discriminate those students with the necessary basic linguistic skills from those who do not have them. Some support for this view is found in the fact that on both easy and medium texts, no difference was found between the grammatically-correct and the acceptable-form-class scoring procedures. Since one would normally expect some difference between these procedures - a non-native might be expected to choose correctly the grammatical functions of subject (credit under the acceptable-form-class scheme) and yet to make mistakes in concord or number (thereby losing a point on the grammatically-correct scheme) - and since there was indeed a difference between the two on the difficult text, one is led to the conclusion that the two easier texts were not adequate tests of the students' grammatical abilities, in that they do not tax the reader adequately, they do not cause him to make grammatical or lexical errors which might be induced by lack of comprehension of the text.

9.1.3.2 Scoring procedures

The main study, apart from showing an interaction between text, deletion rate and scoring procedure, also showed that significant differences in mean scores were achieved by different scoring procedures. In general, the expectation was confirmed that when the criterion for correct replacement takes account of more features of the text, it becomes harder to answer the item correctly. It is easier for non-native speakers to produce a grammatically correct replacement for a cloze blank than it is for them to produce a semantically acceptable response. The most difficult task is to replace the missing word exactly. This result confirms all previous research findings. However, Fillenbaum et al (1963) presumed that different scores would depend upon different features of the environment and predicted that whereas a verbatim (exact) response depends upon both remote and close linguistic constraint, the correct identification of the form class of the blank would depend merely on the immediate context. If such were the case, one would expect different scoring procedures to be affected differently by varying the amount of context available. Since this did not happen - or at least since changes in deletion frequency did not bring about the expected constant change in the different scores - it must be concluded that different types of responses do not depend upon different amounts of context. It can only be assumed that if these scoring procedures do measure different aspects of contextual constraint, the increase in constraint must be linked with the quality of the context rather than its quantity.

Two facts exist which throw doubt on the supposition that

different procedures measure different things. The first fact was also revealed by other studies of different scoring procedures with non-native speakers, namely, that the procedures intercorrelate at a high level. If the procedures were genuinely different, then one would expect considerably lower intercorrelations. The second fact is that the different procedures do not reveal different patterns of relationships, neither with the dictation nor with the various subtests of the ELBA. Nor do the factor analyses show a different factor structure for different scoring procedures. If the different procedures were measuring different aspects of English proficiency, one would expect one procedure to relate more than another to certain of the proficiency measures used. This did not happen, and it is necessary to account for the result.

One explanation is that any cloze score is made up fo a variety of different responses, and that inevitably any scoring procedure is measuring several things at once. It is clearly the case that one scoring procedure is not exclusive to responses of one type. For example, the grammatically-correct procedure does not mark only for grammaticality - since the exact word is by definition grammatical it would be allowed on the grammatically-correct procedure. Thus the argument would run that since any given scoring procedure is not a pure measure, a classification is necessary of the responses to cloze tests. These responses, arranged into mutually exclusive categories like grammatically-correct-only, same-form-class only, semantically-acceptable only, would then be related to external measure of grammatical, lexical and other linguistic abilities, in order to see whether

the making of a particular type of response, rather than the acquisition of a particular type of score, reflected a particular kind of ability. This issue is, of course, insoluble within the framework of the present study.

A different explanation for the apparently close relationship among these scoring procedures is that the cloze task is complex and requires many different abilities. It was suggested that the factor analyses carried out provided only incomplete solutions to the question of what cloze tests. If this is true, then it might be the case that the use of different measures in the equations might result in a more satisfactory analysis of the different scoring procedures. For example, previous studies have shown the importance of intelligence in the performance of the cloze task, at least for native speakers, and perhaps such a variable should be included in the analysis. In this view, then, the different scoring procedures do measure different skills, but unfortunately this study was not delicate enough to show this.

A third possibility is that the scoring procedures all measure the same ability, and that one scoring procedure is as good as any other for such measurement. This explanation seems a little unlikely in view of the fact, not only that the procedures clearly allow or exclude different types of responses, but also that different procedures relate somewhat differently to the criterion tests. Although different procedures do not reveal different patterns of relationships, some procedures simply relate more to measures of English-as-a-foreign-language proficiency than do others. The semantically-acceptable procedure consistently correlates higher than any other scoring procedure with

the ELBA Total, with virtually all of the subtests, and with the dictation. The exact word method usually correlates only third best, after the grammatically-correct procedure.

This fact suggests a fourth possibility, namely, that although the procedures do not measure different skills, different procedures measure different amounts of the same skill. This view is also supported by the fact that different procedures have different amounts of loading on the same factors.

Of course, this solution begs the question as to what the different procedures all measure, and leaves unanswered the question as to why the semantically-acceptable procedure, for example, should correlate more closely with either the dictation or the ELBA than the exact word method. Any such answer must at this stage be speculative, but it will be suggested in greater detail later that the cloze is a measure of relatively low-order linguistic skills. This means that the cloze is less sensitive to discourse and is more sentence-bound than is ordinarily supposed. If that is the case, then a semantically-acceptable procedure which usually rules out constraint from beyond the sentence - i.e., discourse constraint - is a truer reflection of the cloze task than is the exact word procedure. Moreover, such low order linguistic skills might be supposed to be measured by what have been termed "core proficiency" tests (like grammar and vocabulary) rather than by inferential tests. Hence the closer relationship between proficiency in English as a foreign language as measured by ELBA and the semantically-acceptable procedure than the exact word procedure.

Whatever the solution - and the problem is not, and cannot be,

solved here - the fact remains that this study confirms several studies and confounds others in that the semantically-acceptable procedure is seen to be a better measure of English-as-a-foreign-language proficiency than the exact word method. Not only are the correlation coefficients higher for this procedure, but so too are the reliability coefficients (the exact word procedure is the least reliable procedure). The semantically-acceptable procedure also gives a more efficient distribution of item difficulty and discrimination on two of the three texts. Moreover, it reduces the differences between cloze tests due to the use of a different deletion frequency, and so, since the study in Chapter 5 (Section 5.4.3) showed that even non-native speaker judges can make reliable and valid judgments of acceptability, it is to be preferred, for use with non-native speaker subjects, to the exact word procedure.

9.1.3.3 The dictation and cloze

Before proceeding to a discussion of the third experimental variable - deletion frequency - which will lead directly into a discussion of what cloze tests, the relationship between the cloze and dictation will be discussed.

Previous research had suggested that the cloze was more closely related to the dictation than to traditional tests of proficiency in English as a foreign language. It was asserted that this was because both the dictation and the cloze were integrative tests, measuring the learner's internal expectancy grammar. In such a case, the modality of the test was less important than the fact that both were tapping the same psycholinguistic ability. This study sought to verify this, and to attempt to account for it by relating the dictation to different types

of cloze tests, formed by manipulation of three experimental variables. However, previous findings were not confirmed sinze cloze and dictation proved to be no closer than cloze and core proficiency in English as a foreign language. The factor analysis revealed cloze to be more associated, more frequently, with this core proficiency than with a dictation factor. Moreover, the dictation was less associated with the more integrative parts of the ELBA than with discrete-point tests of grammar, vocabulary or sound recognition. The difference between the two dictations was also unexpected, given previous research findings. The easy and difficult dictations did not intercorrelate as high as would be expected if they were both measuring the same thing. Also, the easy dictation related more closely to the ELBA and to the cloze than did the difficult dictation. Although there was no suggestion from the factor analysis that the two dictations were testing different abilities, it is probable that one dictation tests more of whatever abilities the dictation is testing than does the other. Given the closer relationship between the difficult dictation and the difficult cloze, it appears likely that both demand a better command of basic linguistic skills - lexical and syntactic - than do the easy cloze and the easy dictation (both of which are more closely related to each other than to either of the difficult tests). Thus what the dictation and cloze have in common is a measurement of formal linguistic skills, which measurement improves as the difficulty of either test increases.

The fact that a scoring procedure (the key word scheme) which clearly samples only part of the task - that is, it does not measure the task in a truly integrated fashion - has exactly the same correlations

as the more normal procedure suggests that dictation is not best described by the phrase "integrative test".

Can the dictation be said to be a high-level test if a different scoring procedure - measuring only the ability to identify nouns and verbs - gives the same results? If the dictation were a test of higher-order skills, then presumably such a scoring procedure would not be an adequate measure of such skills.

Moreover, the evidence provided by student errors showed that subjects trying to make sense of chunks of language, and trying to make sense of the language inside the chunk, but not in relation to other chunks. Thus creative errors are often syntactically correct within themselves, but semantic or pragmatic nonsense when compared with adjacent chunks.

It seems, then, from this study that the reason dictation has correlated highly with these measures of ability in English as a foreign language is that it, too, is such a measure. The reason it correlates more with some subtests than with others does not appear to be due to the claimed fact that it is an integrative test, but because it is essentially a test of low-level linguistic skills. Hence the dictation correlates best with those cloze tests, texts and scoring methods which themselves best allow the measurement of these skills. The relatively high correlations with other cloze tests can still be seen as due to the same factor, since a large proportion of any cloze test will inevitably call only on low-level linguistic skills. This also accounts for the apparently contradictory findings from previous research that sometimes dictation correlates high with reading tests and sometimes it

correlates low (see Chapter 5). When the reading test involves mainly linguistic skills - fact-finding, answering referential, low-level questions - it will correlate high with dictation; when the test involves more higher-order skills - answering inferential and evaluational questions - dictation will correlate at a lower level. This has nothing to do with the "integration" or "discreteness" of the test. If by integrative tests is meant the measurement of several things at once, the term is trivial, since, inevitably, all tests test more than what they purport to test - obvious examples are the widespread tests of listening which involve reading printed text.

Furthermore, because dictation tests low-level skills, it is, for such a population as was used in this study, basically an easy test (hence the skewed distributions which resulted even from the very difficult dictation, which affects its efficiency and makes it a dubious test for widespread use, despite good reliability figures and good correlations with other measures of English-as-a-foreign-language proficiency.

The study has shown that the variables of deletion frequency, text and scoring procedure all have an effect on the relationship of the cloze with dictation, and that, moreover, the difficulty of the dictation also has an effect. In addition, it has shown that not only does the cloze not necessarily relate more closely to dictation than to other measures of proficiency in English as a foreign language, but also that the dictation frequently relates more to these measures than to cloze.

The monolithic view of cloze's relationship with dictation must now be modified with provisos regarding the effect of the variables. Oller's results (1971, 1974, 1975), showing that cloze relates more to

dictation than to traditional measures of English-as-a-foreign-language proficiency, have not been replicated by this study, which has, indeed, confirmed that cloze has a relatively distant relationship with reading comprehension and that it is often more related to straight core proficiency tests like grammar and vocabulary than to dictation.

9.1.3.4 Deletion frequency

Previous research, quoted in Chapter 3, into the effect of the amount of context on the predictability of missing letters, indicated that beyond 32 letters there was no noticeable increase in constraint up to 10,000 letters. This number of letters was translated into words - somewhere between 4 and 8 - and related to other experimental results which showed that ten words provided maximum constraint, beyond which no increase was discernible, whereas less than four words provided noticeably less constraint. The results of MacGinitie's 1960 study were quoted as showing that no difference was to be expected between cloze deletion frequencies, provided that at least five words intervened between blanks. However, more recent studies of the effect of variation in deletion frequency on the cloze suggest that a different deletion rate might give a different measure of the readability of text. The Algerian study appeared to confirm MacGinitie's finding that a test with less than four words between blanks would always be significantly more difficult than one with more context. Moreover, there was tendency for the tests to get easier as the deletions became less frequent, up to a deletion rate of every tenth word. This result seemed to confirm the previous findings that maximum constraint operates at a level somewhere between 4 and 10 words. The main study found no significant deletion

rate effect in a two-way analysis of variance, which would suggest that MacGinitie's conclusion - that any deletion frequency beyond every fifth word will give the same results - is correct. However, both this result and the Algerian study results, as well as all previous research, ignored the variable of text difficulty. The main study showed that texts were significantly different and had a significant interaction effect. When text differences were taken into account in the main study, significant differences were found for some deletion rates, but not for others. No consistency in difference or in direction of difference was found. Similarly, when individual tests were examined in the Algerian study it was clear that there was no consistency in the differences that were found. If no interaction is apparent between deletion rate and text, then one could conclude that contextual constraint is not affected by differences in text difficulty. Since there was an interaction, however, not only does the deletion frequency become a crucial variable in readability studies, and, more generally, in the construction of a cloze test, it also means that the Algerian study's tentative general conclusions about deletion rates are invalidated, because they ignored the text variable.

This does not mean that different deletion rates test different deletion rates test different abilities - that a frequent deletion rate tests grammatical abilities or that a less frequent deletion rate tests comprehension. The evidence from both the Algerian and the main studies is that although different tests based on different deletion rates relate differently both to a total measure of English-as-a-foreign-language proficiency and to subtests designed to test different skills,

there is, again, no consistency in the differences between deletion rates. Less frequent deletion rates do not tend to correlate more with one ability than another, nor does any one deletion rate, regardless of text, provide a better measure of English-as-a-foreign-language proficiency.

Although significant differences were found amongst tests constructed using different deletion rates, when identical items with varying amounts of context were compared, no significant differences whatsoever were found. In other words, increasing the context around a blank from five words to eleven words has no effect on the difficulty non-native speakers find with that blank. It does not make the missing item more predictable than the provision of only five words did. This is true for every text, and every scoring procedure used. This finding agrees with MacGinitie's conclusion that increasing context beyond five words has no effect. Yet when the effect of varying deletion frequency is examined on a cloze test, rather than on certain test items, significant differences between cloze tests do emerge.

This means that the differences there are between deletion rates on a cloze test are due entirely to the other items chosen as a result of the selection procedure, and not to the length of the context between gaps. Inevitably, a different deletion pattern results in different items being deleted, and this will lead to the different results on the test.

Further, the lack of patterning in deletion rate differences is due entirely to the choice of items to be deleted.

Thus, although increasing the amount of context round a cloze

item beyond five words has no effect, using different deletion rates to produce cloze tests does have an effect. One must distinguish between the manipulation of context for experimental purposes - the former case - and the creation of a cloze test for practical purposes. The findings of both MacGinitie and this study therefore only have theoretical relevance, since it does not follow from the conclusion that amount of context has no effect that therefore the deletion rate chosen has no effect either. In practice, it does matter that Oller uses deletion rate 7, Anderson uses deletion rate 8, Bormuth deletion rate 5, and others deletion rate 10. Even with 50 items in a cloze test, choosing a different deletion rate results in a significantly different test, which can give different measures both of readability and of reading ability. Worse still, this effect, as we have seen, is not predictable, since it depends entirely on which other items are deleted. It remains for others to determine whether an increase in the number of items deleted - perhaps 100 or even 200 - could reduce or remove the effect of changing deletion rates. This study shows, however, that the only way to reduce the effect is to use a different scoring procedure from the exact word method. The grammatically-correct method consistently showed no difference between deletion rates; the same holds for the same-grammatical-function procedure. Even the semantically-acceptable method proved to be capable of reducing the effect of deletion rate.

If cloze is sensitive to deletion rate changes but not to the change in <u>amount</u> of context - i.e., if different words are deleted, different results are achieved, even with 50 items - then cloze is more word-based than was previously thought. There is no evidence that the

nature of the cloze task changes, since validating correlations for different deletion rates do not differ, yet different means result for different cloze deletion rates. This can only mean that cloze scores are very much determined by which words have been deleted (cf., Bormuth's (1964c) finding that even at the same deletion rate, starting deletion at different points results in different cloze scores for over half the twenty passages used, even when 50 items were deleted).

In other words, yet again, as different words are deleted, a different cloze score results. If cloze is so sensitive to the deletion of different words, then one wonders exactly what it is testing, or indeed, whether it is capable of measuring higher-order skills.

A related point is that if cloze items are, on the average, just as easy or as difficult with five words of context as with eleven words, then closure must be based on a very small amount of context indeed. In fact, the basis for closure would seem to be merely the phrase or, at best, clause in which the item is to be found. This provides evidence not only for the claim that cloze is sentence-bound (see Chapter 3), but goes even further, to suggest that cloze is clause- or even phrase-bound. That being so, one would not expect it to be capable of measuring higher-order skills, but rather to be a measure of much lower-order skills, and therefore to be extremely sensitive to selection of different words, as indeed it has been seen to be.

9.1.3.5 What the cloze tests

Indirect support for the contention that cloze is not a good test of reading comprehension comes from previous research into reading gain with native speakers of English. If cloze were a measure of

reading comprehension one would expect scores on the task to increase considerably after reading the unmutilated passage on which the cloze is based. Chapter 1 quoted several studies which indicated that the random cloze was insensitive to such reading gain and so, one might infer, to comprehension of text. It is suggested here that comprehension of a text involves more than understanding the lexis and syntax of each sentence, and thus comprehension would then be considered to be a higher-order skill. The accumulation of evidence in the main study suggests that for non-native speakers the cloze is a test of relatively low-order skills, which are not closely related to reading comprehension as it is ordinarily tested, but which are certainly closely connected with what has been termed "core proficiency". Unlike Oller and Conrad (1971), who found cloze to correlate highest with reading comprehension and lowest with grammar and vocabulary, Darnell (1968) found that cloze correlated lowest with reading comprehension and highest with vocabulary and grammar tests. This study has confirmed Darnell's findings. In particular the factor analysis showed cloze to be only remotely associated with reading comprehension, and much more closely associated with core proficiency (grammar and vocabulary). The factor analysis of the ELBA had shown it to be composed of a core factor, and a metalinguistic, or "higher" factor, which was suggestive of a division into two separate levels of skills - one the ability to handle the syntax and lexis of the language (dubbed "core proficiency"), and the other the ability to make inferences and to handle text. The cloze was never associated with this second factor, but was frequently associated with the former.

The first point, then, is that cloze for non-native speakers

is more a test of low-order linguistic skills than of higher-order inferential abilities. It is possible to take this dichotomy further, using the evidence of the factor analysis of ELBA, and suggest that whilst the higher-order skills involve the ability to handle text, the lower-order skills relate to the ability to handle sentences. This division enables one to relate what is being said about the cloze here to the discussion in Chapter 3 as to whether cloze is sentence-bound. If cloze were sentence-bound, one would not expect it to be capable of measuring skills that govern the ability to handle units larger than the sentence.

Moreover, the main study has produced evidence that the basis for closure is almost certainly a very restricted context, which reinforces the argument that cloze is at best sentence-bound, and supports Carroll's contention (Carroll and Freedle, 1972) that cloze essentially measures local (i.e., phrase and clause) redundancy. Such a conclusion would also account for the fact that different cloze versions result in different mean scores, since sentence-boundness would involve greater dependence on the words of each sentence, and thus greater sensitivity to the deletion of different words than would be the case if the cloze were text-bound. Clearly the fact that the cloze procedure deletes words rather than phrases or clauses must limit its ability to test comprehension of more than the immediate environment, since individual words do not usually carry textual cohesion and discourse coherence (with the obvious exception of cohesive devices like anaphora, lexical repetition and logical connectors). Moreover, the high correlation of the semantically-acceptable scoring procedure with the measure of

English-as-a-foreign-language proficiency, and the high intercorrelation of this procedure with the exact word method (the former a procedure which is designed to be insensitive to long-range contextual constraint), both appear to add support to the thesis that cloze is essentially sentence-bound.

Thus the second point is that the evidence of this study, from analysis of both the deletion frequency and the scoring procedure variables, reinforces the evidence quoted in Chapter 3 which points to the fact that cloze is essentially sentence-bound. This is not to deny that certain cloze items are capable of testing comprehension of more than the immediate environment, but merely to assert that as a test the cloze is largely confined to the immediate environment of a blank. That more might be involved in the cloze test is apparent not only from the results of previous research, which has in some cases suggested that much variance in the cloze test is not accounted for by the abilities tested on the multiple-choice tests with which it is usually compared. (In particular, the role of intelligence cannot be underestimated, at least for native speakers.) This study, too, has suggested that cloze might be factorially more complex than appears to be the case.

Previous writers have suggested that cloze measures three things, all related to some extent: closure, the language correspondence between reader and writer, and linguistic redundancy. Doubt has been cast in the past on the extent to which the cloze measures these areas (see Chapter 1), and a brief reference to these concepts seems necessary in the light of the present study.

Taylor justified the cloze procedure by saying that it re-

quires and involves closure, that is, the human tendency to perceive in conformity with familiar shapes. Given the creative nature of language (as expounded by Chomsky, amongst others), one would not expect a text to have a familiar shape, but rather the sentences, clauses and phrases of which that text is composed. One would expect, for example, transitional probabilities to be greater within a short linguistic unit (phrase or clause) than within a larger unit, and, indeed, research into the statistical structure of text has shown such to be the case. Thus one would a priori expect closure in conformity with familiar shapes to be based upon relatively short stretches of language. The present study has, in fact, confirmed such a supposition by showing that the addition of context has no effect on closure. Thus the present results are by no means incompatible with the proposal that cloze is based on closure, and in fact, support MacGinitie's claim (1971) that cloze is based on familiar patterns of expression and not on comprehension. However, the results do suggest a need for a closer definition of what exactly linguistic closure is, and on what it might be based.

The notion that cloze is a measure of the similarity of the language of reader and writer was also put forward by Taylor in 1953, and has been referred to in order to justify the use of the cloze procedure as a test of proficiency in English as a foreign language. The present study indicates that such a notion might need rethinking. The idea of language correspondence between reader and writer requires that replacing the exact word used by the writer represent the highest degree of agreement between the two parties. By extension, replacing the exact word should provide the best measure of the correspondence of the linguistic

system of the reader/test-taker and the target language. However, the evidence shows that the best measure of proficiency in English as a foreign language is achieved not by scoring correct only the exact word deleted, but by allowing as correct any replacement which is semantically acceptable. Similarly, the finding of the deletion frequency investigation does not accord with what would be expected from the notion of language correspondence. Presumably, the more language there is available for evidence of the linguistic system of the writer, the more likely it is that a reader can conform to this system. This does not happen, however. Perhaps the view needs to be modified to refer simply to the language correspondence, over short chunks of language, between writer and reader. This would then allow the semantically-acceptable procedure to be the best scoring procedure, since it rules out reference to longer chunks of language.

The notion that cloze is a measure of redundancy has already been criticised by Bowers and Nacke (1971), who point out (see Chapter 1) that linguistic redundancy is not the same as transitional probability, and does not operate through the accumulation of words, or rather, that redundancy is not governed by the number of words available. If their view of redundancy is accepted, then the results of this study do not contradict the assertion that cloze measures redundancy, since merely increasing the amount of context does not increase the cloze score, whereas a traditional view of redundancy would expect cloze scores to increase with increasing context. However, if linguistic redundancy is not based on the statistical structure of text, but upon its syntactic, semantic and pragmatic structure, then a random cloze test is probably in principle not a suitable measure of such redundancy.

In summary, it appears that for non-native speakers of English the cloze test is not notably an integrative test, nor a test of reading comprehension and high-order skills, nor of the ability to handle text rather than sentences, but that it is more a sentence-bound test of low-order linguistic skills closely related to core proficiency tests of English as a foreign language.

9.2 The cloze test with native speakers of English

The native speaker study investigated the same variables text, deletion rate and scoring procedure - examined in the non-native speaker study, but did not relate the cloze tests, nor, therefore, any of the experimental variables, to external criteria.

9.2.1 Text

The texts were ranked in the same order, regardless of deletion rate or scoring procedure. It proved easier, for example, to supply a grammatically correct answer on an easy text than on a medium text. Admittedly, the texts were intended to be obviously different, and it could be that more similar texts would not be unfailingly distinguished whatever the deletion rate or the scoring procedure, but these results show any cloze to provide a consistent measure of readability. However, there was significant interaction between text and deletion rate, which could mean that with other less different texts, using a different deletion rate might result in a different measure of readability.

9.2.2 Scoring procedures

On the whole, different scoring procedures on the same test produce different mean scores. This is always true on a difficult text,

usually true on a medium text, but only sometimes true on an easy text. However, the lack of difference between scoring procedures on an easy text is due to the subjects' having achieved virtually maximum scores. When maximum scores are not achieved, the procedures usually produce significant different mean scores.

The order of difficulty was that the exact word procedure was always most difficult, and the grammatically-correct procedure easiest. The semantically-acceptable was easier than the exact word, but harder than the identical-form-class procedure.

This result shows both that as more context is allowed to constrain a blank, the more difficult becomes the restoration of the deleted word, and that native speakers may not always guess the form class of the deletion, but they will tend to supply an answer which is grammatically correct.

Contrary to expectations, and contrary to the results of the non-native speaker study, the intercorrelations of the scoring procedures were quite low, with only 20% achieving a level of .9 or more. Although there was relatively high agreement among the grammatical scoring procedures (GRCO, IDFC and ACFC), the exact word method showed very little relationship to the grammatically-correct procedure (GRCO). This result suggests that, if the exact word method is a measure of reading comprehension, then the ability to fill a blank grammatically is not related to the ability to comprehend text. Moreover, the exact procedure was not very closely related to the semantically-acceptable procedure (correlations of the order of .7 to .8) and this relationship varied according to the text and deletion rate used. This relationship is lower than previous research suggested, and also

lower than that experienced in the non-native speaker study.

Finally, it might be expected that different scoring procedures require different amounts of context for success - possibly, a grammatical procedure requires only three or four words of context, whereas the exact word procedure requires ten or eleven words. However, this did not prove to be the case, since different scoring procedures did not react in a predictable manner to a change in deletion frequency.

9.2.3 Deletion rate

Significant differences were found for the exact word on all texts, but the semantically-acceptable procedure only showed significant differences on the easy text, and most grammatical procedures failed to reveal any significant differences among deletion rates. It is possible that the grammatical procedures tended to hide differences in deletion rates because they always achieved high mean scores with these native speakers. To reduce the effect of deletion rate on a cloze test with native speakers, therefore, it would appear sensible to use any scoring procedure other than the exact word method.

The direction of such differences as were found was not as expected. One would predict a steady increase in comprehension as the quantity of the context increased. In fact, however, this only occurred on the medium text. On the easy text, the reverse was true, i.e., that as deletions became less frequent, the test became more difficult, not less. On the difficult text, the middle two deletion rates (8 and 10) proved to be more difficult than either 6 or 12. The conclusion, thus, must be that what differences between deletion rates do occur are not consistent or predictable.

When identical items were compared across deletion rates, no significant differences were found. In other words, increasing the context of a cloze blank from five to eleven words does not increase the predictability of that blank. This is the same finding as that made on the non-native speaker study.

9.2.4 Cloze as a test

As on the non-native speaker study, the problem was encountered of how to evaluate the cloze as a test. Again, traditional tools like measures of central tendency, of dispersion, of reliability, item difficulty and item discrimination were used to compare cloze tests, if only because they permit comparisons with other testing techniques. Their adequacy was far from apparent, however.

In general, the cloze tests seemed to be fairly inefficient, from the point of view of item analysis, with many extremely easy and difficult items, regardless of text or scoring procedure. Rarely were more than 60% of the items within acceptable limits of difficulty. When item discrimination was added, there were never more than 40% of items acceptable on both counts. On average, a mere 14 out of 50 items were acceptable, which indicates a high level of inefficiency.

The exact word scoring procedure produced the best distribution of item difficulty and discrimination, but only on the easy and medium texts. On the difficult text, the semantically-acceptable procedure was more efficient.

The grammatical scoring procedures, especially the grammatically-correct procedure, were the least efficient scoring procedures from the point of view, not only of the item analysis, but also of

measures of central tendency and dispersion: they resulted in high means with little dispersion. Again, the exact word method resulted in the most satisfactory means and distributions.

The conclusion is that if one intends to score cloze by the exact word method, it would be preferable, in traditional terms, to use a relatively easy or medium text with native speakers. If a difficult text is to be used, then the test would best be scored by the semantically-acceptable scoring procedure.

9.3 A comparison of native and non-native speaker performance

The most striking thing to emerge from the summary of this study is the similarity of results for both native and non-native speakers of English. In all essential respects, the results are identical. The three texts were ranked in the same order of difficulty by both groups of subjects, and in both cases, the texts were distinguished one from the other regardless of deletion frequency or scoring procedure. Nevertheless, in both cases a significant interaction was found between text and deletion rate, which indicates that in native speaker studies as well as in non-native speaker studies, the text must be taken into account when doing research into the deletion frequency, and the deletion frequency must be taken into account in readability research.

Similar findings were made in the deletion frequency study, namely that for both populations, changing the deletion rate may result in a significantly different test, but that this change is neither consistent nor predictable. Interestingly, the direction of difference was identical for native and non-native speakers, but in neither case did it accord with expectations. For native speakers as for non-native speakers,

the use of a scoring procedure different from the exact word method reduced or removed the difference in deletion rate, and for both groups of subjects, no differences were found when only identical items were contrasted.

In both cases it was concluded that differences in cloze tests are not due to the change in deletion frequency from one test to another, but simply to the fact that different words are deleted by different deletion rates. Just as in the case of the non-native speakers, native speakers are not aided in their ability to restore deleted words by the addition, even the doubling, of context around the deletion.

Different scoring procedures had similar effects in both cases; the same rank order of difficulty caused by the procedures was seen for native speakers as for non-native speakers. The most difficult procedure was the exact word method, followed by the semantically-acceptable procedure, whilst the easiest, for native speakers and non-native speakers, was the grammatically-correct procedure.

The examination of test efficiency also produced similar findings with regard to score distributions, item difficulty and item discrimination, as affected by the experimental variables. For native as for non-native speakers it was concluded that the best test was produced by the exact word method on an easy text, or by the semantically-acceptable procedure on a difficult text.

The only real difference to emerge between the two populations was the finding that for native speakers, the different scoring procedures intercorrelated at a much lower level than for non-native speakers. Although the same trend was seen in the non-native speaker

study - namely that the grammatical procedures relate closely to each other, but that the exact word procedure shows little relationship to the grammatically-correct procedure - it was much more pronounced in the native speaker study. Unfortunately, the lack of external criteria in the native speaker study means that it is not possible to determine whether the low intercorrelations of scoring procedures reflect real differences in what they measure. It is possible to speculate that the exact word procedure is a better measure of reading comprehension abilities for native speakers than for non-native speakers, hence its lower correlation with the other scoring procedures, but it is not clear that this is so, or why it should be. Possibly native speakers' greater experience of and ease with the bones of the language - the lexis and syntax - enables them to see distant relationships among ideas in text more easily than non-native speakers, who might be held up to a greater extent by difficulties with the language at sentence level. One would then expect the exact cloze to be less a test of lower-order skills than it is for non-native speakers. But if this were so, then one would expect changes in deletion frequency to have less effect for native speakers than for non-native speakers. However, such changes actually had equivalent effects. A suggestion that the nature of the cloze is different for native and non-native speakers would also predict that if for non-native speakers the cloze is sentence-bound, for native speakers it is not. Yet changing the amount of context around identical items had no effect in either case. In fact, the findings of this study, that in all essential respects manipulating experimental variables in the cloze is the same for native as for non-native speakers, indicate that

in all essential respects the cloze task is the same for native and non-native speakers, and that therefore the tentative conclusions reached for non-native speakers - that cloze is a test of low-order skills, that it is essentially sentence-bound, and so on - also apply to native speakers. Of course, it is impossible to conclude that cloze is a measure of low-order core proficiency for native speakers, but the results available do suggest that it is sentence-bound. This, in turn, predicts that the cloze is limited to testing a lower level of skill, which one would not term core proficiency in English as a foreign language, but which would be related to such proficiency inasmuch as it refers to the ability to handle the lexis and syntax of the language adequately.

Of course, one expects native speakers to be able to handle such linguistic tasks with perfect ease. Nevertheless, there are several indications in these results that such is not the case. Firstly, although the native speakers can do the grammatical cloze, the scores are not perfect, and indicate a considerable number of grammatical errors on the part of the subjects. This in part may be due to incorrect chunking of text, or to subjects ignoring (or not noticing) grammatical constraint from outside the immediate environment of the blank. Secondly, the results clearly show a great similarity in the performance of native and non-native speakers. This has been outlined above.

However, a series of direct comparisons has yet to be made of the performance of native and non-native speakers on each cloze test. Table 9.1 sets out the mean scores for both groups on each cloze test scored by all five procedures. t-tests on the results show native

speakers to have been virtually always significantly superior in performance to non-native speakers. The only exceptions to this are on the easy text, usually with (relatively easy) form class scoring procedures. However, the differences in mean scores are not very great. In particular, grammatical scores differ, on the medium and easy texts, by only 2 to 14 percentage points. (Incidentally, the surprisingly small differences between native and non-native speakers in grammatical scores on the easy text (2 to 6 percentage points) support the suggestion that the easy text does not tax the non-native speakers' command of low-order linguistic skills, especially since on the difficult text there are much greater differences between the two groups on the grammatical scores). Yet even the exact word method produces differences between the two groups of only ten pencentage points on the difficult text.

In fact, the semantically-acceptable procedure consistently produces the greatest differences between native and non-native speakers, of the order of eight to twenty-six percentage points (further evidence for the superiority of the semantically-acceptable procedure over the exact word method, if separation of groups is what one requires of a cloze test). Even the grammatically-correct procedure produces greater differences than the exact procedure on the difficult and medium texts. In general, the greatest differences between native and non-native speakers appear on the difficult text, but even here the maximum difference achieved is twenty-six percentage points.

The results of this contrast of the performance of the two groups indicate that the bulk of one population - the native speaker - is higher up the scale of achievement than is the other (i.e., that on

the whole native speakers are better at cloze than non-native speakers, and especially at finding semantically-acceptable replacements). Nevertheless, considerable overlap in performance between the two groups is seen. Table 9.2 sets out, for three scoring procedures, the amount of this overlap. From these figures it can be seen that the cloze test never achieves complete separation of native from non-native speakers. The amount of overlap varies greatly, from 27% to 97%, but virtually half of the tests produce over 80% overlap between natives and non-natives. Clearly, the distribution of scores for non-native speakers is greater than that of native speakers, which means the performance of non-native speakers is more variable, or that the native speakers form a more homogeneous group. Yet occasionally one or two native speakers actually perform worse on the cloze than the worst non-native speaker (e.g., E8, E10, N12), and frequently non-native as well as native speakers achieve maximum scores. (Again, interestingly, the least overlap between groups is attained by the semantically-acceptable procedure on the difficult text.)

These results confirm findings from previous research. Carroll et al (1959) found not only that presumed improvement in foreign language achievement was not measured by cloze since it could not distinguish first-year students from second- or third-year, but also that the cloze could not distinguish native from non-native speakers of a language. Similarly, Oller and Conrad (1971) found that cloze could not distinguish advanced learners of English as a foreign language from native speakers of English in the freshman year of college. Indeed, Carroll et al felt that cloze was not a good measure of foreign

language proficiency because it was necessary to allow for the subjects' ability to do the cloze in their native language. Certainly, it is commonly supposed that a good foreign language proficiency test will measure abilities that native speakers possess, and that therefore a clear separation will be achieved of native from non-native speakers. This study shows that although significant differences are gained between the two groups, clear separation is not attained, since there was considerable overlap of performance on all cloze tests. If it is held that separation by proficiency test is a sine qua non of such a test, then the cloze test appears to be an inadequate proficiency test.

However, it is possible that, especially in reading tests, native speakers do not differ as greatly from non-native speakers as they are imagined to in tests of listening and speaking, since other cognitive variables might be imagined to intervene in the reading process (including, possibly, the intelligence variable said to be important in cloze tasks). If it is the case that proficiency in reading a foreign language is not greatly different from proficiency in reading in one's mother tongue, then one might expect more overlap in performance of native and non-native speakers than a naïve view of foreign language proficiency would predict.

Although it is not possible to resolve this issue within the framework of this study, it is an uncontrovertible fact that the performances of native and non-native speakers on these cloze tests are very similar, and that no clear separation of the two groups has been achieved.

The non-native study clearly showed that the cloze was related

not to reading comprehension at a high level, but to core proficiency or to lower-order skills. There is no reason to suppose that the native speaker study produced different results, although, because of the lack of external criteria, a definitive answer cannot be given. The supposition, then, is that the cloze is a low-order test, in nature sentence-bound, measuring local redundancy, familiarity with certain patterns of expression, closure based on short chunks of language, rather than measuring discourse constraints, or the ability to relate ideas in a text to one another or to evaluate the information and ideas in text, for native as well as for non-native speakers. Thus, the claim that the cloze is not a test of reading comprehension applies equally to native and non-native speakers.

If this is true, then it would appear that native and non-native speakers of English do not differ greatly from each other in their command of low-order skills, at least when these skills are tested in writing. (Such written tests obviously involve a basic ability to read, which is not the reading comprehension referred to in this study, but which may very well condition the native speaker's ability to respond appropriately to the cloze task.) In other words, what the cloze task measures varies not only in the non-native speaker population, but also within the native speaker population. It may be that what is being measured is the ability to take cloze tests, as Carroll et al claim, but it would seem necessary to define this ability more closely. If the interpretation of the results of this study is correct, this ability is basically linguistic in nature, and is of a relatively low cognitive order. The implications of this conclusion for the testing of foreign

language proficiency are considerable.

CHAPTER 10

Conclusions and Implications

10.1 Conclusions

The following were the null forms of the hypotheses tested in the main study:

1a. There is no significant difference between cloze scores when deletion frequency changes.

1b. There is no significant interaction between deletion rate and text for easy, medium and difficult texts.

2. There is no difference in ranking of texts by different deletion rates.

Subhypothesis: There is no difference in ranking of texts by different deletion rates when scored by different methods.

3a. There is no significant difference between exact and other scoring methods.

3b. There is no significant difference between deletion rates when scored by exact word, and when scored by other methods.

4a. There is no difference between deletion rates as measures of proficiency in English as a foreign language.

4b. There is no difference between texts as measures of proficiency in English as a foreign language.

4c. There is no difference between scoring methods as measures of proficiency in English as a foreign language.

4d. There is no interaction between deletion rate, text and scoring

method as predictors of proficiency in English as a foreign language.

The null forms of Hypothesis 2 and the subhypothesis were accepted, but the null forms were rejected for the remaining hypotheses, that is, 1, 3 and 4. The following, then, are the results of the study, with reference to the experimental hypotheses:

Hypothesis 1 a) Significant differences between cloze score are found when deletion frequency changes. However, there is no significant difference between deletion frequencies when identical items are compared. This is true for native and non-native speakers.

b) There is a significant interaction between deletion rate and text for easy, medium and difficult texts. This is true for native and non-native speakers.

Hypothesis 2 There is no difference in ranking of (obviously different) texts by different deletion rates.

Subhypothesis There is no difference in ranking of texts by different deletion rates when scored by different methods. Hypothesis 2 and the subhypothesis are true for native and non-native speakers of English.

Hypothesis 3 a) Significant differences exist between exact and other scoring methods.

b) Significant differences exist between deletion rates when scored by the exact word, and when scored by other methods. Other scoring procedures reduce the

number of significant differences. Sections a) and b) are true for both native and non-native speakers of English.

Hypothesis 4 a) There are differences between deletion rates as measures of English-as-a-foreign-language proficiency.

b) There are differences between texts as measure of English-as-a-foreign-language proficiency.

c) There are differences between scoring methods as measures of English-as-a-foreign-language proficiency. The differences in 4a) and c) are not, however, predictable, since

d) There is an interaction between deletion rate, text and scoring method as predictors of English-as-a-foreign-language proficiency.

It is possible, from these results, to draw a series of conclusions. The first one is that the effect of the experimental variables is essentially the same for native speakers of English as it is for non-native speakers; therefore, most of the following conclusions apply equally to both populations.

The second major conclusion is that the cloze procedure is not a unitary procedure, since there is a marked lack of comparability among the tests it may be used to produce. The fact emerges very clearly from this study that different cloze tests, produced by variations in certain of the variables, give unpredictably different measures, particularly of proficiency in English as a foreign language, but also, probably, of other abilities and of readability.

Nevertheless, it can also be concluded that changes in at least two of the variables do not result in the cloze test measuring different abilities. It is not the case that a more frequent deletion of words will necessarily result in a different ability being measured, nor that scoring the cloze by a different procedure will mean that different skills are being tapped. Different scoring procedures do not measure different abilities, any more than do different deletion frequencies, despite the fact that a change in either scoring procedure or deletion frequency can result in both significantly different scores and different correlations with external criteria, on any given text.

The inescapable conclusion from the study of deletion frequency is that neither for non-native speakers of English, nor for native speakers, is an increase in predictability gained by a context of eleven words rather than five words. One might expect that the form class of a deletion is just as predictable with five words of context as with nine words, but the results show clearly that it is just as easy, or as difficult, to restore the exact word deleted with contextual clues of five words as with twice as much context. It is unlikely that the amount of context is going to have any effect if increased beyond eleven words; therefore, if quantity of context is to have any effect, it must be at a level lower than five words.

The second, equally important conclusion, if somewhat obvious from a common sense point of view, is that what makes one cloze test different from another is the fact that different words are chosen for deletion. This follows from the fact that although items common to two deletion rates result in similar scores (i.e., regardless of the amount

of context) tests using different deletion rates on the same text were found to result in scores which were at least sometimes significantly different from each other.

Therefore, the common assumption is false that the deletion rate used to construct a cloze test is irrelevant. Because one deletion rate deletes different words from a passage from those deleted by another deletion rate, a different picture is gained of the readability of the text or the reading comprehension of those subjects from the picture that might have been gained if a different deletion rate had been used. This, at least, is applicable for cloze tests of 50 items. The basic problem, again, is that of the non-comparability of cloze tests.

The quantity of context has no effect on the predictability of deleted words, beyond a minimum level. Poor students are not helped in their ability to restore words by an increase in context. It is not easier to find a grammatically correct replacement with more context, nor to find a semantically acceptable word. Nor is it necessarily easier to do a cloze task on a difficult text with more context. Of course, one consequence of these facts is that contextual constraint may be seen to vary according to the quality of the context rather than its quantity. However, they also imply that enough information is provided by five words of context, or, at least, that more information will not be provided by an increase in context. For this reason, among others, it was suggested that the cloze is essentially sentence-bound.

It is further concluded that the cloze is related at least as much to supposedly discrete-point tests as to integrative tests, and,

in particular, that it relates more to traditional tests of core proficiency than to tests like the dictation. It is a better test of ability to deal with syntax and lexis at sentence level than of reading comprehension in general, the ability to handle metalanguages, or of inferential or deductive abilities; in short, of what have here been termed higher-order abilities.

Easy texts seemed to be a less adequate test of this core proficiency than were more difficult texts, but no evidence was found to support a hypothesis that, in contrast to difficult texts, easy texts permit the measurement of reading comprehension or global skills. It would appear that easy texts also measure low-order skills, but that they do not measure them as well as more difficult texts.

Although no scoring procedure measured any different ability, the semantically-acceptable procedure appeared to be superior to any other, including the exact word method, because it correlated best with criterion measures of proficiency, improved the differentiation achieved by the cloze between native and non-native speakers of English, reduced the effects of the variables of text and deletion rate on the prediction of proficiency in English as a foreign language, and also reduced the the differences in mean scores of different deletion rates. It resulted in improved score distributions on both medium and difficult texts, improved reliability figures, improved item facility and discrimination statistics, and a reduced incidence of extreme scores.

It must further be concluded that the simplicity of the cloze procedure is misleading, and that it does not automatically result in a good test, at least when measured by traditional testing standards. The

results showed that the test user needs to analyse the cloze test just as carefully as he would examine the results of a more traditional test of the multiple-choice type. In general, the procedure seemed to result in inefficient tests with, on average, only 28% of items performing satisfactorily.

Finally, in relation to the measurement of readability, it was concluded that the cloze procedure can distinguish obviously different texts regardless of methodological variations. Nevertheless, if it can only distinguish such texts, it would appear to be of limited use, since such texts can probably be distinguished more easily by, for example, teacher judgments.

10.2 Implications

Some of the implications of these conclusions are obvious, some perhaps less so. The implications for the use of the cloze procedure in testing are manifest. The procedure must be treated as any other technique, and generalisations about what it tests have to be made very cautiously, if at all (just as one would normally refrain from making statements about what multiple-choice tests measure). Testers should above all be aware that changing the deletion rate or the scoring procedure or using a different text, may very well result in a radically different test, and not give them the measure that they expect. However, they should also be aware of its limitations, if it is true that the cloze test is essentially a sentence-bound test of lower-order skills. If the tester uses the technique, there are certain methodological recommendations that can be made on the basis of this study. One is to delete at most every fifth word, but in order to reduce the effect of

the deletion frequency chosen, to score by the semantically-acceptable procedure rather than by the exact word method. (As outlined above, there are also other advantages in using such a scoring procedure.) Another is that to gain a good measure of core proficiency, a relatively difficult text should be used if possible.

Another implication of these research findings is that current assumptions, based on some published research, about the usefulness of the cloze may well be wrong. It would seem that the widespread use of the cloze as an automatically valid test of something needs discouraging in order for further reflection and experimentation to take place. Of course, when experimentation is reported, the report should automatically include full details of the deletion frequency used, the type of text and the scoring procedure.

One problem of the analysis of the cloze as a test is what one is to do about poor results when they are discovered. If an item analysis approach is adopted, what are the consequences of the identification of poor items? Manipulating items, or ignoring them in the scoring clearly destroys the principle of randomness of the initial selection of items. However, if randomness is <u>not</u> an essential feature of the cloze as a test of proficiency in English as a foreign language or whatever, as opposed to its use as a measure of readability, then perhaps the tester would be better advised, instead of abandoning the principle when revising the test <u>after</u> analysis, to abandon it <u>before</u> constructing the test. In other words, perhaps the test constructor should use a rational cloze, selecting items for deletion based upon what is known about language, about difficulty in text, and about the

way the language words in a particular text.

Of course, Taylor's original point about the desirability of randomness in the deletion of words from text was that it would provide an adequate and representative sample of text difficulty, and that by selecting items for deletion, text difficulty would be distorted. It would, for example, be possible to make an easy text look very difficult by deleting the few unpredictable words and leaving the rest. There is, however, no reason why this principle should be applied to the testing of the reading abilities of subjects. There still remains the question, nevertheless, of whether random cloze really provides a suitable measure of the readability of text. Sampling all the words or a representative selection of a text may still not give a true picture of text difficulty, which conceivably is not expressed in individual words, but rather through syntactic complexity, abstractness of ideas, inexplicitness of connections, "poor writing", etc., which are not capturable by the cloze, not only because it is sentence-bound, or a measure of lower-order skills, but also simply because it deletes individual words. Thus, although one solution might be to use rational rather than random deletions, using all possible linguistic insights, another solution might be to delete linguistic units that do not necessarily confine themselves to the orthographic word. A procedure could be imagined, for example, that deleted selected morphemes, or noun phrases, or all modification, rather than "words". Concepts could be deleted, which would not necessarily be expressed in specific lexical items.

The tentative conclusion that cloze is sentence-bound and a measure of local redundancy and lower-order skills carries implications

for the traditional view of the nature of the cloze. If cloze is a measure of the correspondence of the language systems of reader and writer, then this can only be true over short chunks of language, since the cloze could not be a measure of the total linguistic correspondence. If cloze involves closure, then, similarly, that closure must be based on little context, and possibly on transitional probabilities, rather than on the whole text. The cloze does not appear to measure redundancy as traditionally defined, since amount of context has little effect, yet the random cloze is in principle incapable of measuring linguistic redundancy as defined by Bowers and Nacke (Chapter 1). To be such a measure, it would be necessary to use the rational cloze, at least.

In fact, this study has been of the random cloze, and reference to "cloze" has frequently been synonymous with reference to random cloze, largely because it is the random cloze which is the most widespread technique. One of the major implications of this study, however, is that the emphasis on random selection be downgraded, and that the rational deletion of items be given more consideration, and be subject to further research.

What is the implication of the finding that native and non-native speakers of English perform in a very similar manner on the cloze? One possible implication is that testers should not necessarily expect native speakers to do well on a foreign language proficiency test. Therefore, clear separation of native from non-native speakers need not be demanded of such a test. However, this removes one of the easiest and most commonly used methods of validating a foreign language test (by trying it out on native speakers). Another, different, implication is

that because the cloze does not separate the two populations, it is not a suitable measure of foreign language proficiency, that, instead,it is a measure of some extraneous cognitive variable common to all language users.

10.3 Areas for further research

Such implications clearly demand research. This study has very clearly pointed up the need for further research into the relationship between one's mother tongue and one's second language. What is the influence of the ability, or lack of it, to infer, to deduce, to evaluate, to see relationships in text, and so on, in one's mother tongue, on one's ability to read in a second language? Does the ability to read one's first language influence the success one will have in reading a foreign language? Do other cognitive variables have an influence both on the use of the mother tongue, and on the foreign language?

Further research is clearly also called for into the nature of the so-called lower-order skills, and their relationships between languages. Can it be that difficulties with foreign language syntax and lexis are traceable to similar difficulties in one's mother tongue? Are these linguistic divisions - syntax and lexis - justifiable in terms of what the cloze test demands of subjects?

In any case, there is a clear need to devise ways of unambiguously testing both lower-order and higher-order skills. Once a clearly higher-order reading comprehension test, and a clearly lower-order comprehension test have been devised, then the cloze should be compared with both, to test the hypothesis that it would relate most closely to the lower-order skills test.

It might be possible to develop ways of making the cloze test into a measure of higher-order abilities. Two such ways have already been suggested - viz., using a rational deletion scheme rather than a random one, or deleting linguistic units which would not necessarily be confined to single words. Such suggestions would need extensive empirical validation.

Of particular relevance to the recommendation that work be done on the rational cloze is the suggestion made in Chapter 9 that a definition is needed of linguistic closure - what exactly it is, how it works, and why. Research is needed into the nature of the familiarity of a given linguistic pattern or environment. Even more, research is required into the nature of contextual constraint and predictability, in order to see what determines the correct replacement of deletions, and to provide a firm basis for rational deletion. One possible way of doing this would be to carry out an extended error analysis on cloze tests; in other words, to analyse responses made by subjects to investigate the way in which they violate or comply with constraint - to see, for instance, whether they show the influence of distant or proximate constraint. If it were possible to classify responses in terms of their violation of near and far constraint, they could then be correlated with validated measures of low- and high-order skills to see whether making responses that do not violate long-range constraint depends on the presence of high-order skills, and conversely, whether the lack of such skills leads to an inability to make such responses.

This suggestion is related to that made in Chapter 9 (section 9.1.3.2) that a future project might classify cloze responses as

grammatically-correct only, semantically-acceptable only, and so on, and then relate production of such responses to criterion measures. One might then seek to answer questions like: Is it the case that a subject who responds only with grammatically-correct responses - but not with semantically-acceptable responses - is less advanced in English as a foreign language and therefore that his performance would correlate higher with a grammar test than with a reading comprehension test, whereas someone who can respond with the exact word deleted is at least as good on the reading test as on the grammar test? However, even this would need to be tested differently according to the linguistic category of the deleted item, since finding the exact replacement for a function word is easier (since this is a closed set) than for a content word. Indeed, it might be that the random cloze is entirely unsuitable for this sort of exercise, and that a rational cloze deletion scheme would be more meaningful.

A further area for investigation might well be to compare the advantages of rational versus random cloze as measures of English-as-a-foreign-language proficiency, and to examine whether they measure different underlying abilities.

Yet another point to emerge from the study is the need for a closer investigation of the nature of proficiency in English as a foreign language, and particularly of core proficiency. What are the skills that go into this proficiency, and why is general listening comprehension so closely associated with reading tests of grammar and vocabulary?

Such an investigation might also usefully relate to native speaker linguistic abilities.

A clear need has been seen for a further study of the dictation, to ascertain what the nature of the skills it is testing might be. One possible way would be to look at different scoring procedures; another might be to undertake a rigorous and thorough error analysis.

One drawback of the present study has been that the ELBA and dictation tests used in the investigation were themselves factorially complex tests, despite the simplicity suggested by their titles, and this did not facilitate the identification of emergent factors. Enough evidence has been gathered to indicate that further study may be fruitful. A subsequent investigation into what cloze tests for non-native speakers should perhaps include tests of psychological abilities whose underlying factors are more or less well defined, and which are thought, for theoretical reasons, to be relevant to the cloze. Such tests might include a measure of closure, a measure of inferential reading abilities, and a measure of purely sentence-processing abilities, as well as some standardised English-as-a-foreign-language proficiency measure.

At a more concrete level, several projects suggest themselves directly from the results of this study. The fact that cloze can discriminate obviously different texts, regardless of the changes in the variables studied, has been established. What is now needed is a replication which would examine texts of greater similarity to see if varying certain variables still had no effect. As already suggested, the cloze has little value as a measure of readability for non-native speakers if it only distinguishes clearly different texts.

The necessity has become obvious for a test analysis technique that could be used more appropriately on tests like the cloze. Research

is needed into the development of such a tool, which would hopefully avoid the assumptions of a traditional item analysis, and relate more to the presumed nature of the cloze.

Finally, replication is desirable of this study of deletion frequency with longer tests, perhaps up to 200 items, to see whether increased length removes the differences found between deletion rates.

Whatever research may be carried out into the cloze procedure in the future, whatever the true nature of the cloze may be, whatever use may be made of the technique in language testing or language teaching, this study has shown that it would be wise to bear in mind the remarks of Rankin quoted in Chapter 1:

> "Performance on a cloze test . . . is influenced by the reading ability of the reader and the difficulty of the materials (and) . . . the type and number of items deleted. Until we know more about the possible interrelationships of these variables . . . we should be cautious in interpreting cloze tests."