Journal of Research in Reading *Vol 2 No. 2 1979*

The effect on the cloze test of changes in deletion frequency

J. Charles Alderson, *University of Michigan.*

ABSTRACT

Although the pseudo-random cloze procedure has been in use for some twenty-five years as a measure of readability and reading comprehension, little research has been carried out into the effect of deleting words from text more or less frequently. This paper reports on an experiment in which the deletion frequency variable was systematically studied. Every 6th, 8th, 10th and 12th word was removed from three texts of differing difficulty, and the effect studied. Significant differences among cloze tests resulted, but the differences were unpredictable. Deleting every 12th word did not necessarily result in an easier test than deleting every 6th, 8th or 10th word. However, when only items identical to both cloze tests under consideration were compared, no significant differences were found. It appears that cloze items are, on the whole, unaffected by context greater than five words. Testers are warned that changing deletion frequency may result in a different measure of readability or comprehension.

RÉSUMÉ

L'effet sur le test de Cloze des changements de la fréquence de l'élimination des mots

Bien qu'on ait utilisé le procédé de Cloze pseudo-hasardeux comme mesure de lisibilité et de compréhension en lecture pendant quelque vingt-cinq ans, on a fait peu de recherches sur l'effet qu'a l'élimination d'un texte d'un nombre de mots plus ou moins important. Après une revue des recherches, y compris des expériences de l'effet qu'ont des quantités diverses de contexte sur la capacité de prédire des lettres et des mots, le rapport rend compte d'une expérience où on a examiné les fréquences variables de l'élimination des mots. On a barré un mot sur six, sur dix et sur douze dans trois textes de difficulté diverse pour obtenir douze tests de Cloze. On a administré ces douze tests de Cloze à des élèves d'un collège d'enseignement secondaire en Écosse et à des élèves qui faisaient des études au Royaume-Uni dont la langue maternelle n'était pas l'anglais; on a marqué les résultats d'après cinq procédés différents. Des différences significatives se sont manifestées parmi les tests de Cloze, mais les différences étaient contradictoires et impossibles à prédire. La suppression d'un mot sur douze n'a pas nécessairement donné un test plus facile que la suppression d'un mot sur six, sur huit ou sur dix. Les résultats étaient essentiellement les mêmes pour les Anglais que pour les autres. Quand on a comparé les résultats acquis sur le test de Cloze par les élèves étrangers avec les résultats acquis sur un test standard de Compétence en anglais langue étrangère, on a trouvé que les corrélations variaient considérablement d'un test de Cloze à l'autre.

Le changement dans la fréquence de la suppression des mots employée pour produire un test de Cloze a donné non seulement des degrés divers de lisibilité mais aussi de compétence linguistique. Il est suggéré que les expérimentateurs veillent à ce que les tests produits par le procédé de Cloze soient rendus valables. Mais, comme modifier le test après l'analyse des détails et après des épreuves de validité ne peut qu'aller à l'encontre du caractère pseudo-hasardeux de la sélection initiale, il serait peut-être préférable d'abandonner le principe de la suppression et de sélectionner les rubriques à l'avance selon un principe linguistique ou psycholinguistique.

De plus, on a étudié les détails qui se trouvaient communs aux deux tests de Cloze et on a trouvé que dans l'ensemble l'importance du contexte dans la rubrique de Cloze ne produit pas d'effet sur la capacité à prédire cette rubrique. Bien que ce résultat suggère que le test de Cloze peut rester insensible à un contexte éloigné, cela ne signifie pas que le nombre de suppressions opérées pour produire un test de Cloze n'a rien à voir avec cette question-ci.

INTRODUCTION

For some twenty years the cloze procedure has been used to construct tests purporting to measure the readability of text and the reading comprehension abilities of subjects, initially with native speakers of English and latterly also with non-native speakers of that language. The most usual form of the procedure has been the pseudo-random removal of words from text by deleting every nth word, where n is usually somewhere between five and ten. The use of a pseudo-random deletion was originally justified (Taylor 1953) by reference to the purpose of the test, which was to measure text readability. It was claimed that a (pseudo-) random sampling of words in text would provide a more adequate and valid characterisation of text difficulty than would the deletion of words according to some subjective principle. Clearly, it is possible to remove words from an otherwise easy text which would be difficult to restore, and equally possible to remove easily restorable words from a difficult text. In both cases, the biased deletion would result in a distorted picture of the text's difficulty.

This justification would not necessarily hold for the measurement of reading comprehension, however: the need for an adequate sample of the words in a given text is not obvious. Nevertheless, the practice in cloze test construction has been to use the pseudo-random deletion procedure to produce cloze tests of 'reading comprehension'.

Relatively little research has been done into the effect of changing the deletion frequency on the pseudo-random cloze procedure, and actual practice has assumed that the frequency has little effect, provided that at least four words of context appear between deletions. This practice has been partly based on the results of research by MacGinitie (1960) which seemed to show that whereas there was a significant difference between deletion rate

3 (deleting every 3rd word) and deletion rates 6, 12 and 24, there was no difference between deletion rate 6 and 12, or 6 and 24, or 12 and 24. MacGinitie attempted to account for his results by suggesting that the redundancy in English 'for restorative purposes' acts mainly with small segments of speech, and that 'the units in which thoughts are composed may seldom be greater than five or six words'.

It should be pointed out that MacGinitie's findings ignored the differences between texts, and were based only on those words which were deleted in all his tests.

Research on the effect of amount of context on restorability began with the information theorists, who used the Shannon guessing game (Shannon 1951), in which subjects guess which letter comes next in a series of letters (and therefore, words). Burton and Licklider (1955) found that the constraint imposed by preceding context of 32 letters was little less than that imposed by 10,000 letters, although considerably greater than that imposed by 1, 2, 4 and 8 preceding letters. Shepard (1963) found that a context of 40 words did not impose significantly more constraint than a context of 10 words on the number of words subjects could restore to a deletion in a given amount of time. Aborn, Rubinstein and Sterling (1959) found that a context of less than 4 words between deletions substantially reduced constraint whereas increasing context between deletions beyond ten words did not increase subjects' abilities to restore the deletion. This finding they related to Burton and Licklider's, by suggesting that 32 letters represent between 4 and 8 words. However, it is somewhat difficult to relate this to the cloze procedure since they used isolated sentences rather than text.

Salzinger, Portnoy and Feldman (1962) found no difference between deletion rates 5 and 7 on passages representing different orders of statistical approximation to English and thus 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'. However, Fillenbaum *et al* (1963) did find differences between deletion rates 5 and 6.

Miller and French (1974), using deletion rates 5, 7 and 10, found deletion rate 7 to be easier than the other two frequencies, but were unable to account for this, whilst McNinch *et al* (1974) found no consistently easiest or most difficult deletion rate. Their results showed that varying deletion patterns significantly affects the measurement of readability but the lack of consistency makes it impossible to generalise from their results to other texts. Their results are, however, interesting in that they lead one to question the assumption that deletion frequency has no effect on cloze scores provided that words are not deleted more frequently than every fifth word, and they encourage speculation that different deletion patterns might produce different results.

The only study of the effect of deletion frequency on non-native speakers of English (Haskell 1973) found no significant differences between passages'

mean scores for deletion rates 5, 7, and 10. No studies have been made of the effect of changing deletion frequency on validating correlations with comprehension or linguistic proficiency criteria, nor has any direct comparison been made between the differential performances of native and non-native speakers of English. Above all, no attempt has been made to account for those research findings that show no differences among deletion rates and those that do show differences.

The Study

In order to investigate the effect of changing the deletion frequency on a cloze test, three texts were chosen to represent high, medium and low levels of difficulty respectively from the content area of fictional writing. Fiction was chosen so as not to bias difficulty in favour of the particular subject experience of any one group of readers. The levels of difficulty were determined both by readability formulae (Dale-Chall, Flesch, Fog and Smog) and by pooling the judgements of 19 experienced reading teachers. For further details, see Alderson (1978). From each text every 6th, 8th, 10th and 12th word was removed to give twelve cloze tests in all. The responses were scored by five different procedures: the exact word procedure, a procedure which allowed any semantically acceptable word (SEMAC), one which allowed any grammatically correct word (GRCO) and two form-class procedures, one allowing as correct any restoration from the same form class as the deletion (IDFC) and the other allowing restorations from an acceptable form class which filled the same grammatical function as the deleted word (ACFC). The cloze tests were distributed randomly to 360 native speakers of English (secondary pupils aged 15-16, in the Edinburgh area, all of whom were judged by their teachers to be at least moderately capable readers) and 360 non-native speakers of English who were in the UK going through tertiary education, aged 18 and over. Thus, 30 native speakers and 30 non-native speakers responded to any given cloze test. In addition, most of the non-native speakers also took two dictation tests (one difficult, one easy) and the *English Language Battery* by Elisabeth Ingram, University of Edinburgh. This latter battery consists of seven subtests: Sound Recognition (1), Intonation (2), Stress (3), Listening Comprehension (4), Grammar (5), Vocabulary (6) and Reading Comprehension (7).

RESULTS

Table 1 sets out the mean scores for each cloze test, scored by the five different procedures, with native speakers of English, and Table 2 gives similar results for the non-native speakers.

TEXT LEVEL		SCORING PROCEDURE				
		EXACT	SEMAC	GRCO	IDFC	ACFC
Difficult Text	Do6	19.6	33.2	43.9	35.0	37.4
	Do8	15.9	34.5	44.4	36.5	38.9
	D10	14.7	31.8	41.2	36.9	38.8
	D12	20.3	34.9	43.3	38.2	39.6
Medium Text	Mo6	25.5	38.5	45.6	43.3	44.8
	Mo8	24.9	39.5	46.6	41.3	44.5
	M10	29.8	41.7	46.5	44.9	46.3
	M12	29.0	39.1	44.1	43.1	44.0
Easy Text	Eo6	34.3	46.0	48.4	45.6	46.7
	Eo8	34.9	45.3	48.4	45.4	47.1
	E10	32.6	43.5	47.2	46.5	46.6
	E12	30.1	43.4	47.5	44.6	46.5

Table 1: *Mean scores for cloze tests: native speakers of English.*

TEXT LEVEL		SCORING PROCEDURE				
		EXACT	SEMAC	GRCO	IDFC	ACFC
Difficult Text	Do6	14.8	24.9	37.8	31.2	33.0
	Do8	10.0	21.6	31.3	25.6	27.8
	D10	9.4	19.0	30.7	27.0	28.7
	D12	14.6	24.7	35.7	30.1	31.4
Medium Text	Mo6	20.1	30.3	39.1	38.9	39.8
	Mo8	19.9	31.3	40.4	37.5	41.2
	M10	23.4	31.6	38.8	39.5	41.2
	M12	21.0	28.5	36.6	37.8	38.5
Easy Text	Eo6	30.3	42.4	45.9	44.3	45.7
	Eo8	29.8	40.7	44.6	42.7	44.7
	E10	30.8	40.3	44.2	44.6	44.7
	E12	26.7	38.7	44.4	42.1	44.5

Table 2: *Mean scores for cloze tests: non-native speakers of English.*

These results show that changing the deletion rate has an apparent effect on the mean score. On the medium text, deletion rates 10 and 12 are notably easier than the other two deletion rates when scored by the exact word procedure (Table 1). The results show that the differences between

mean scores are not always as expected. Common-sense suggests that as the deletions become less frequent, words are easier to replace. This would mean that a test based on a deletion frequency of every sixth word would be more difficult than a test based on the deletion of every eighth word, which would be more difficult than a test constructed by deleting every tenth word, and so on. Figures 1 and 2 set out the results for the exact word scoring procedure.

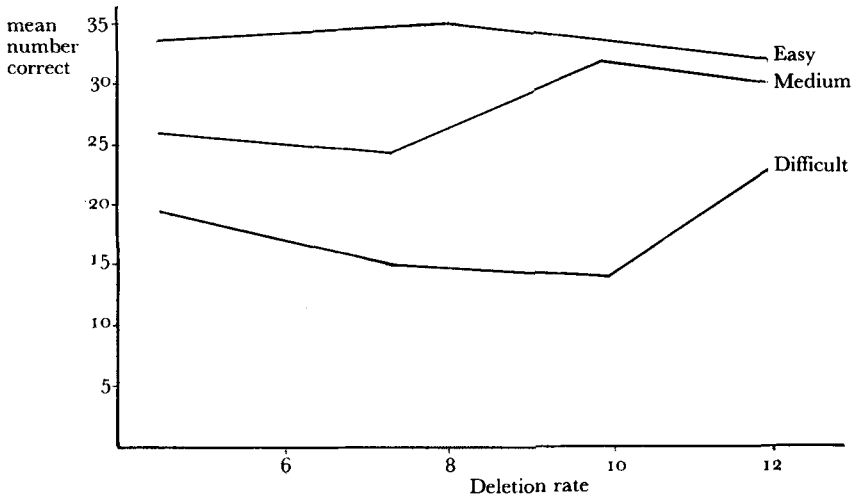

Figure 1: *Mean scores for 3 levels of text difficulty using the exact word procedure: native speakers of English.*

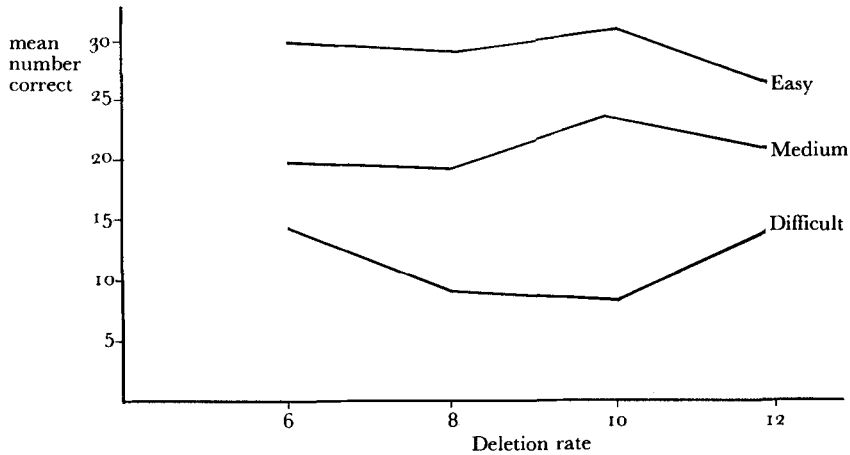

Figure 2: *Mean scores for 3 levels of text difficulty using the exact word procedure: non-native speakers of English.*

These figures show the commonsense supposition to be erroneous since, for example, on the difficult text deletion rates 6 and 12 were not different from each other, but both were easier than deletion rates 8 and 10.

One-way analyses of variance were carried out to test for the significance of the differences between mean scores, and where the analysis of variance revealed significant differences, t-tests were calculated to test for significant differences between pairs of means. The analysis of variance for native speakers found no significant differences between means for the GRCO and ACFC scoring procedures, on the IDFC procedure no differences were found on the difficult and easy texts, and on the SEMAC procedure none were found on the medium and difficult text (Table 3).

Text level	SCORING PROCEDURE				
	EXACT	SEMAC	GRCO	IDFC	ACFC
Difficult	sig	NS	NS	NS	NS
Medium	sig	NS	NS	sig	NS
Easy	sig	sig	NS	NS	NS
sig = significant at 5% level NS = not significant					

Table 3: *Results of Analysis of Variance on deletion rates: native speakers of English.*

For non-native speakers, the results were similar (Table 4) in that approximately one third of the tests showed significant differences. In this case, however, the medium text never resulted in significant differences among deletion rates.

Text level	SCORING PROCEDURES				
	EXACT	SEMAC	GRCO	IDFC	ACFC
Difficult	sig	sig	sig	sig	NS
Medium	NS	NS	NS	NS	NS
Easy	sig	NS	NS	sig	NS
sig = significant at 5% level NS = not significant					

Table 4: *Results of Analysis of Variance on deletion rates: non-native speakers of English.*

The results of the t-tests for significant differences between pairs of means are set out in tables 5 and 6 below.

Deletion rates	Text level and scoring procedure				
	Difficult EXACT	Medium EXACT	Medium IDFC	Easy EXACT	Easy SEMAC
6 : 8	sig	NS	sig	NS	NS
6 : 10	sig	sig	NS	NS	sig
6 : 12	NS	sig	NS	sig	sig
8 : 10	NS	sig	sig	NS	NS
8 : 12	sig	sig	NS	sig	NS
10 : 12	sig	NS	NS	sig	NS

Table 5: *Differences between mean scores of deletion rates, native speakers of English.*

Deletion rates	Text level and scoring procedure					
	Difficult EXACT	Difficult SEMAC	Difficult GRCO	Difficult IDFC	Easy EXACT	Easy IDFC
6 : 8	sig	NS	sig	sig	NS	NS
6 : 10	sig	sig	sig	sig	NS	NS
6 : 12	NS	NS	NS	NS	*sig*	*sig*
8 : 10	NS	NS	NS	NS	NS	NS
8 : 12	sig	NS	NS	sig	NS	NS
10 : 12	sig	sig	sig	NS	sig	sig

Table 6: *Differences between mean scores of deletion rates, non-native speakers of English.*

No consistent pattern emerges from these results. Significant differences do exist among deletion rates, but they are not always in the expected direction and the differences vary from text to text and from scoring procedure to scoring procedure. The conclusion would appear to be that using a different deletion frequency to produce a cloze test may result in significant differences but the differences are neither consistent nor predictable. However, using any scoring procedure other than the strictest (the exact word) reduces and often removes the differences between tests due to the use of different deletion rates.

However, 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 in 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.

Only those items common to both pairs of any comparison were selected and t-tests calculated for the differences between the means of these identical items in those cases (see tables 5 and 6) where differences had been found between the cloze tests. It was assumed that where no differences had been found, differences between deletion rates for identical items would not exist.

Deletion rates	Text level and scoring procedure				
	Difficult EXACT	Medium EXACT	Medium IDFC	Easy EXACT	Easy SEMAC
6 : 8	NS	$p<.05$	$p<.05$	NS	NS
6 : 10	NS	NS	NS	NS	NS
6 : 12	NS	NS	NS	NS	NS
8 : 10	NS	NS	NS	NS	NS
8 : 12	NS	NS	$p<.05$	NS	NS
10 : 12	NS	NS	$p<.05$	NS	NS

Table 7: *Differences between deletion rates for identical items, native speakers of English.*

Deletion rates	Text level and scoring procedure					
	Difficult EXACT	Difficult SEMAC	Difficult GRCO	Difficult IDFC	Easy EXACT	Easy IDFC
6 : 8	NS	NS	NS	NS	NS	NS
6 : 10	NS	NS	NS	NS	NS	NS
6 : 12	NS	NS	NS	NS	NS	NS
8 : 10	$p<.05$	NS	NS	NS	NS	NS
8 : 12	NS	NS	NS	NS	NS	NS
10 : 12	NS	NS	NS	NS	NS	NS

Table 8: *Differences between deletion rates for identical items, non-native speakers of English.*

From these results (Tables 7 and 8) it is apparent that if non-identical items are excluded from the comparisons, virtually no differences in deletion rates are to be found, and this is true whether one scores by the exact word, the semantically acceptable word or the identical form class procedures.

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. 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.

Since the non-native speakers had also been tested for proficiency in English as a Foreign Language, it is possible to compare cloze tests produced by different deletion frequencies as measures of such proficiency. Table 9 sets out the correlations of the cloze tests scored by five different procedures with the total score of the *English Language Battery* (ELBA) proficiency test.

Text level	Deletion rate	Scoring procedure				
		EXACT	SEMAC	GRCO	IDFC	ACFC
Difficult	6	.51	.67	NS	.43	.43
	8	.82	.87	.73	.80	.74
	10	.79	.83	.79	.83	.82
	12	.77	.85	.68	.72	.70
Medium	6	.86	.88	.81	.67	.68
	8	.68	.77	.74	.51	.50
	10	.57	.74	.75	.70	.65
	12	.73	.78	.75	.70	.69
Easy	6	.59	.74	.60	.44	.45
	8	.70	.69	.61	.50	.46
	10	.65	.74	.75	.63	.65
	12	.67	.77	.72	.73	.71

Table 9: *Correlation of cloze test scores with ELBA total score.*

These results show quite considerable variations in the correlation of the cloze test with a measure of proficiency as the deletion frequency changes. In some cases the relationship is low, in others it is higher. On the difficult text, deleting every 8th word results in a higher correlation than deleting every 6th word (exact word score) whereas on the medium text the opposite is the case. The variation would appear to be somewhat lower when the SEMAC is used. Nevertheless, it is apparent that changing the deletion frequency of a

cloze test changes the validity of that test (at least when used as a test of proficiency in English as a Foreign Language). As with the changes in mean scores produced by varying the deletion frequency, the effect on the measurement of linguistic proficiency is inconsistent and unpredictable. It is not the case that a more frequent deletion will necessarily give a better measure of proficiency, or that a less frequent deletion will consistently prove to produce a more valid test.

CONCLUSION

This study has confirmed MacGinitie's finding that the amount of context between cloze gaps does not have any significant effect on the predictability of the deleted word, providing that at least five words of context are available. This does not mean, however, that providing that at least every 6th word is deleted from the text there will be no effect of changing deletion frequency on the cloze test. When only items identical to both tests are considered the deletion frequency has no effect, but when all the items are considered, i.e., when the two *tests* are compared and not just some items, then the deletion frequency does have an effect. Changing the deletion frequency has an effect not only on the mean score, which means that the measurement of readability is affected, but also on the correlation with external criteria, which means that the validity of the test is affected. Thus, changing the deletion frequency of a cloze test will give a different measure both of the properties of the text and of the abilities of the reader. Moreover, the differences in measurement are entirely unpredictable. It is, therefore, impossible to recommend use of one deletion frequency rather than another since on one text it may result in a higher estimate of readability or a higher correlation with a criterion measure than another deletion frequency, whereas on another text the same deletion frequency may result in a lower estimate or correlation.

Thus, testers, researchers and teachers should not regard the cloze test as automatically valid. It makes no sense to talk of 'the cloze test' or 'the cloze procedure' as measuring X or Y. A specific cloze test may well measure X, but another test produced by the same procedure, using a different text or deletion frequency may well not result in the same measurement of X. This remark would be regarded as obvious if it referred to the multiple-choice technique. 'The multiple-choice test measures reading comprehension' is an absurd remark when it does not refer to a specific test, yet there has been a regrettable tendency in recent years to make precisely that sort of statement about the cloze procedure. The procedure is merely another technique for producing a test, which then needs to be analysed, validated and modified in the usual way. If it is discovered, however, after analysis, that the cloze test needs modification, it is not at all clear how this can be done without affecting the pseudo-random pre-selection of items. If the modification involves selecting some items and discarding others, then one has, *post-hoc*, contravened the principle of the pseudo-random procedure. That being the case, one needs to ask: why use a pseudo-random procedure in the first place?

REFERENCES

ABORN, M., RUBINSTEIN, H. and STERLING, T. D. (1959) Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology, 57* (3), 171-180.

ALDERSON, J. Charles (1978) A study of the cloze procedure with native and non-native speakers of English. Unpublished Ph.D. thesis; University of Edinburgh.

BURTON, N. G. and LICKLIDER, J. C. R. (1955) Long-range constraints in the statistical structure of printed English. *American Journal of Psychology, 68,* 650-653.

FILLENBAUM, S., JONES, L. V. and RAPOPORT, A. (1963) The predictability of words and their grammatical classes as a function of the rate of deletion from a speech transcript. *Journal of Verbal Learning and Verbal Behaviour, 2,* 186-194.

HASKELL, J. F. (1973) Refining the cloze testing and scoring procedures for use with ESL students. Unpublished Ed.D. thesis: Columbia University.

MACGINITIE, W. H. (1960) Contextual constraint in English prose. Unpublished Ph.D. thesis: Columbia University.

MCNINCH, G., KAZELSKIS, R., and COX, J. A. (1974) Appropriate cloze deletion schemes for determining suitability of college textbooks. In P. L. Nacke (ed.) *Interaction: Research and Practice for College-Adult Reading* (23rd Yearbook of the National Reading Conference.) Clemson, S. Carolina: NRC, 249-253.

MILLER, W. D. and FRENCH, S. (1974) Using the cloze procedure to determine the suitability of social science and science textbooks. In P. L. Nacke (ed.) *Interaction: Research and Practice for College-Adult Reading* (23rd Yearbook of the National Reading Conference.) Clemson, S. Carolina: NRC, 254-258.

SALZINGER, K., PORTNOY, S. and FELDMAN, R. S. (1962) The effect of order of approximation to the statistical structure of English on the emission of verbal responses. *Journal of Experimental Psychology, 64,* 52-57.

SHANNON, C. E. (1951) Prediction and entropy of printed English. *Bell System Technical Journal, 30,* 50-64.

SHEPARD, R. N. (1963) Production of constrained associates and the informational uncertainty of the constraint. *American Journal of Psychology, 76* (2), 218-228.

TAYLOR, W. L. (1953) Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30,* 415-433.