TESOL QUARTERLY
Vol. 16, No. 1
March 1982

*The Trait Structure of Cloze Test Scores**

Lyle F. Bachman

Although there is considerable evidence supporting the predictive validity of cloze tests, recent research into the construct validity of cloze tests has produced differing results. Chihara et al. (1977) concluded that cloze tests are sensitive to discourse constraints across sentences, while Alderson (1979) concluded that cloze tests measure only lower-order skills. Anderson (1980) has concluded that cloze tests measure sensitivity to both cohesive relationships and sentence-level syntax. Factor analytic studies (Weaver and Kingston 1963, Ohnmacht et al. 1970) have identified several factors in cloze and other language tests and suggest that cloze deletions should be based on the linguistic and coherence structures of language.

In the present study, the trait structure of a cloze test was examined using confirmatory factor analysis. A cloze passage with rationally selected deletions of syntactic and cohesive items was constructed and given to two groups of non-native English speaking students entering the University of Illinois. A trait structure with three specific traits and one general trait provided the best explanation of the data. The results suggest that a modified cloze passage, using rational deletions, is capable of measuring both syntactic and discourse level relationships in a text, and that this advantage may outweigh considerations of reduced redundancy which underlie random deletion procedures.

There is now a considerable body of research providing sound evidence for the predictive validity of cloze test scores. Cloze tests have been found to be highly correlated with virtually every other type of language test, and with tests of nearly every language skill and component. While it may be reassuring to find such strong relationships between the cloze and other measures of language abilities, if one is interested only in prediction, these relationships are problematical for those concerned with what language abilities the cloze test measures. Indeed, recent research into the construct validity of cloze tests has produced widely differing conclusions. Comparing the performance of non-native English speakers on a standard cloze passage and a scrambled version of the same passage. Chihara et al. (1977) concluded that cloze tests are sensitive to discourse constraints across sentences. Alderson (1979), on the basis of results ob-

* Presented at the 1981 TESOL Midwest Regional Conference and Illinois TESOL/BE Convention, Champaign-Urbana, April 3-4. I would like to acknowledge the assistance of Fred Davidson in the preparation of the data from this research.

Mr. Bachman is Assistant Professor of English as a Second Language at the University of Illinois at Urbana-Champaign. His interests include the development and validation of tests of communicative competence and language learning research methodology.

tained by systematically varying passage difficulty, scoring criteria and deletion ratio, concluded that cloze tests are capable of measuring only lower-order core proficiency skills. Using a rational deletion procedure, Anderson (1980) concluded that cloze tests are capable of measuring sensitivity to cohesive relationships across sentences, as well as sentence-level grammatical structure. In comparing the performance of native and non-native English speakers on cloze tests, Alderson (1980) found considerable variation in native speaker performance, which suggests either that native speakers vary in their ability on lower-level language skills, or that cloze tests measure higher-order skills. Oller and Conrad (1971) interpret a similar finding as an indication that native speakers do vary in "the ability to negotiate language," which presumably includes higher-level skills, and conclude that the cloze is a useful measure of this ability.

The methodology employed in much of this research has been to vary specific aspects of the test—deletion ratio, sequence of sentences, passage difficulty and scoring criteria—and then examine the relationships among these variables, either with the same group of subjects, or across groups of different subjects. Few studies have approached the problem with specific hypotheses of what the cloze measures, made deletions on the basis of these hypotheses, and then examined whether or not the pattern of responses corresponded to the hypothesized predictions. In one such study, Weaver and Kingston (1963) examined the correlations among cloze passages with two types of deletions—random and only main verbs and nouns—and tests of other language skills. Using a principal components analysis, they identified three factors, one of which, a "redundancy utilization" factor, they felt underlies cloze tests. In a similar study, Ohnmacht, Weaver and Kohler (1970) examined the relationships among four different deletion procedures and tests of visual closure, associational fluency, and verbal comprehension. From their principal components analysis they identified closure factors of speed and flexibility, in addition to a well-defined verbal factor. They concluded that researchers should determine cloze deletion patterns according to "language operations and rational operations which are implicit in verbal activity" (pp. 215-216).

In the present study, hypotheses based on a description of semantic relationships in discourse (Halliday and Hassan 1976) were used as a basis for determining the specific deletions. The extent to which this model provided an explanation for the pattern of responses to these deletions was then examined, using confirmatory factor analysis.

1. Procedures

A 365-word passage from an introductory level textbook in social psychology was selected for its appropriateness of difficulty level and representativeness of content. In order to test hypotheses regarding the level of language context measured by cloze tests, three types of deletions

were made: 1) syntactic, which depended only on clause-level context, 2) cohesive, which depended upon the interclausal or intersentential cohesive context, and 3) strategic, which depended on parallel patterns of coherence. These deletions provided 11 syntactic items, 15 cohesive and 4 strategic, for a total of 30 items, with an average deletion ratio of 1:12. An acceptable alternative scoring procedure was used, with the key based on the responses of a native speaker pre-test group and on the acceptable alternatives provided by Hassan and Halliday's description of cohesion. The test passage and the key are presented in Appendix A.

The subjects were two groups of non-native English speaking students who entered the University of Illinois in the Fall term, 1980 ($N = 316$) and the Spring term, 1981 ($N = 102$). They were from a wide variety of language backgrounds and ranged in age from 17 to 47, with a median age of 25. The cloze test was given as part of a placement test battery at the beginning of the semester. This battery also included a 100 item multiple-choice test of structure and a 130-word dictation test. Students were given 20 minutes to complete the cloze test.

In order to avoid analytic problems associated with binary data matrices, the 30 items were grouped to form 13 sets (4 syntactic, 7 cohesive, 2 strategic) according to the similarity of item content.[1] Composite scores were derived by averaging the item scores in each set. The facility indices of these composites ranged from .35 to .77, with a median of .56. The product-moment correlations among these 13 composite scores provided the matrix for the confirmatory factor analyses.

Confirmatory factor analysis is a technique for testing various hypotheses about the causal structure underlying the relationships observed among a number of variables (Jöreskog 1969 and 1978). In using confirmatory factor analysis, the researcher posits one or more factors, each representing a hypothetical trait. This hypothetical factor structure then constitutes a model for predicting the relationships observed in the data. The degree to which the predicted relationships correspond to the observed relationships (correlations) indicates whether or not the model provides an acceptable explanation for the data.

2. Results

The means, standard deviations, and reliabilities for the two groups are given in Table 1.

[1] The intercorrelations among dichotomous variables, such as right-wrong (1,0) item scores frequently yield matrices in which the correlations decrease away from the diagonal. Such simplex matrices are problematical for factor analysis in that their rank is, to a large extent, a function of the number of distinct item difficulty levels. The factor analysis of such matrics, therefore, yields factors which are likely to be interpretable only as difficulty factors. The use of composite item scores minimizes this problem, as well as permitting the use of product-moment correlations rather than tetrachoric correlations, which frequently yield singular matrices. For further discussion of these problems, see Carroll 1945 and 1961, Horst 1965, and Lord and Novick 1968.

TABLE 1

Means, Standard Deviations, and Reliabilities for Cloze Test

Group	N	$\bar{X}$	S	Range	KR_{21}	α
Fall 1980	316	16.69	5.55	0–30	.785	.828
Spring 1981	102	16.83	4.50	6–26	.658	.742

The product-moment correlations among the 13 composite scores for the two groups are given in Appendix B.

Of the numerous factor models which might be posited to explain the correlations among the composite scores, three were of particular theoretical interest. A single general factor which accounts for most of the variance in language test batteries has been reported in many studies in which the cloze test has been used (Stump 1978, Oller 1979, Oller and Hinofotis 1980, Scholz et al. 1980). Although Alderson does not refer to the notion of a general factor, the assumption that cloze tests measure a rather homogeneous group of lower-order core proficiency skills also seems to underlie much of his research. In this study, therefore, one of the models examined posited a single factor underlying all 13 composite scores. A second model posited three independent factors, representing the "completely divisible" competence hypothesis. The third model is one which has found support in numerous studies (Carroll 1967 and 1975, Bachman and Palmer 1981a and 1981b), and posits a general factor plus three specific trait factors.

These three models can be represented schematically as in Figures 1-3 below.

FIGURE 1

General Trait Model

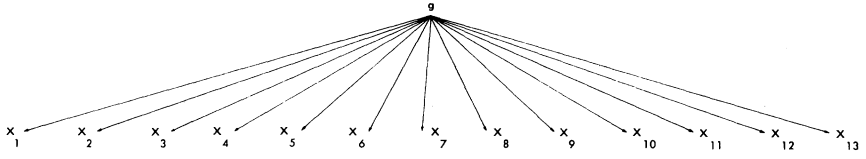


FIGURE 2

Specific Trait Model

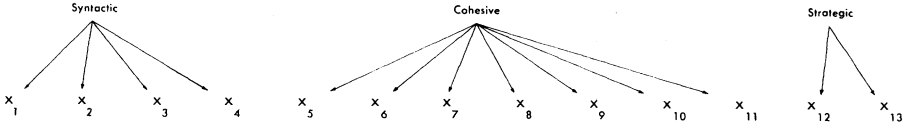
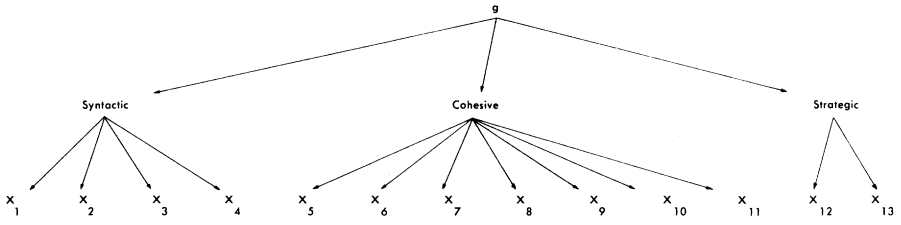


FIGURE 3

General Plus Specific Trait Model



In comparing the extent to which these three models fit the data from the Fall 1980 group, it was found that neither the general trait nor the specific trait model provided a significantly good fit. A model with a general trait and three uncorrelated specific traits, however, did provide a significantly good fit to the data ($\chi^2 = 55.529$, $df = 55$, $p = .7926$, $\Delta = .932$).[2]

An examination of the factor loadings for the general plus specific traits model provides an indication of the relative importance of the factors to each score. Factor loadings for this model are given in Table 2.

TABLE 2

Factor Loadings for General plus Specific Traits Model

	General	Syntactic	Cohesive	Strategic
STX 1	.558	.176	*	*
STX 2	.569	.029	*	*
STX 3	.713	.379	*	*
STX 4	.254	.185	*	*
COH 1	.296	*	.123	*
COH 2	.401	*	.795	*
COH 3	.453	*	.048	*
COH 4	.609	*	.076	*
COH 5	.347	*	.308	*
COH 6	.564	*	.083	*
COH 7	.435	*	.091	*
STG 1	.621	*	*	.525
STG 2	.530	*	*	.186

* (Fixed parameter = 0)

With exception of one cohesive composite score (COH 2), the composites load most heavily on the general factor, with lesser loadings on specific trait factors. Thus, although a model with specific traits provides the best explanation for the data, the effect of a general factor is quite evident.

[2] In confirmatory factor analysis, the logic of hypothesis testing is reversed, so that a small chi square (χ^2), relative to its degrees of freedom (df) corresponds to a high probability (p) of accepting the experimental hypothesis, i.e., the model. The incremental fit index, Δ, provides an indication of the practical significance of the model, in terms of the proportion of the data accounted for by the model (Bentler and Bonett 1980).

While no particular pattern of loadings is apparent, it is interesting that the three highest general factor loadings are on a syntactic, a strategic, and a cohesive composite.

Having found a model which fits the data from the Fall 1980 group, it was of interest to determine to what extent these results were sample dependent. To this end, the general plus specific trait model was tested with the Spring 1981 group, using procedures outlined by Jöreskog (1971). It was found that the model also provided a significantly good fit for this body of data ($\chi^2 = 124.498$, $df = 130$, $p = .6197$, $\Delta = .879$). As is true generally for cross-validation, this fit was slightly less good than with the original group. That it is still a significantly good fit, however, provides additional support for the model.

3. Discussion

The results of this study suggest that cloze tests are not necessarily monotonic tests of general and specific traits. Furthermore, support is found for the claim that cloze tests can be used to measure higher order skills—cohesion and coherence—if a rational deletion procedure is followed.

One possible reason for the inconsistent results of previous research may be the adherence to the principle of random deletion. Alderson has noted that differences in cloze test results may not be due to differences in deletion frequency, but rather to differences in the particular words deleted (1980:72). Random deletion ignores the syntactic and semantic relationships in a text, and is therefore likely to yield inconsistent results, depending upon what proportion of syntactic and textual functions are tapped. If it can be assumed that within a given text more words function syntactically than cohesively, a random deletion procedure would tend to sample a larger proportion of clause-bound words, and therefore appear to be measuring only lower-level skills.

With regard to scoring procedures, semantically acceptable criteria would appear to include cohesive constraints, and it is therefore not surprising that the greatest differences observed between native and non-native speakers have been with this scoring method (Alderson 1980). This is consistent with Anderson's (1980) finding that high level proficiency non-native English speakers performed very similarly to native English speakers on both syntactic and cohesive cloze items, while low level proficiency non-native English speakers performed significantly poorer on cohesive items.

In conclusion, it would appear that cloze passages using a rational deletion procedure can be used to measure textual relationships beyond clause boundaries. The advantages thus gained may well offset considerations for measuring random redundancy or general proficiency.[3] In order to address this latter question directly, however, it will be necessary to

[3] As Alderson has noted, however, such rational deletions must have a theoretical basis (1979:26).

examine the internal structure of cloze passages with random deletions. Identifying the types of deletions which occur in the passages of the numerous studies which have used random deletion, and analyzing the patterns of responses to these deletions through factor analytic procedures (both exploratory and confirmatory) would appear to be a promising starting point for addressing this question.

REFERENCES

Alderson, J. C. 1979. The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly 13*, 2:219-23.

Alderson, J. C. 1980. Native and non-native speaker performance on cloze tests. *Language Learning 30*, 1:59-76.

Anderson, D. C. 1980. Cohesion and the cloze test. Unpublished M.A. thesis, University of Illinois.

Bachman, L. F. and A. S. Palmer. 1981a. The construct validation of the FSI oral interview. *Language Learning 31*, 1.

Bachman, L. F. and A. S. Palmer. 1981b. The construct validation of some tests of communicative competence. Paper presented at 1981 TESOL Convention, Detroit.

Bentler, P. M. and D. G. Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin 88*, 3:588-606.

Carroll, J. B. 1945. The effect of difficulty and clause success on correlations between items or between tests. *Psychometrika 10*, 1-19.

Carroll, J. B. 1961. The nature of data, or how to choose a correlation coefficient. *Psychometrika, 26*, 4:347-372.

Carroll, J. B. 1967. The foreign language attainments of language majors in the senior year: A survey conducted in U.S. colleges and universities. Cambridge, MA: Harvard Graduate School of Education.

Carroll, J. B. 1975. *The teaching of French as a foreign language in eight countries*. New York: Wiley (Halsted).

Chihara, T., J. Oller, K. Weaver, and M. Chavez-Oller. 1977. Are cloze items sensitive to constraints across sentences? *Language Learning 27*, 1:63-73.

Halliday, M. A. K. and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Horst, P. 1965. *Factor analysis of data matrices*. New York: Holt, Rinehart and Winston.

Jöreskog, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika 34*, 183-202.

Jöreskog, K. G. 1971. Simultaneous factor analysis in several populations. *Psychometrika 36*, 4:409-426.

Jöreskog, K. G. 1978. Structural analysis of covariance and correlation matrices. *Psychometrika 43*, 4:443-477.

Lord, F. M. and M. R. Novick. 1968. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Ohnmacht, F. W., W. W. Weaver, and E. T. Kohler. 1970. Cloze and closure: A factorial study. *The Journal of Psychology 74*, 205-217.

Oller, J. W. 1979. *Language tests at school*. London: Longman.

Oller, J. W. and C. A. Conrad. 1971. The cloze techniques and ESL proficiency. *Language Learning 21*, 183-195.

Oller, J. W. & F. B. Hinofotis. 1980. Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In Oller & Perkins (Eds.), *Research in language testing*. Rowley, MA: Newbury House.

Scholz, G., D. Hendricks, R. Spurling, M. Johnson, and L. Vandenberg. 1980. Is language ability divisible or unitary? : A factor analysis of 22 English language proficiency tests. In J. W. Oller and K. Perkins, (Eds.), *Research in language testing*. Rowley, MA: Newbury House.

Stump, T. A. 1978. Cloze and dictation tasks as predictors of intelligence and achievement scores. In J. W. Oller and K. Perkins (Eds.), *Language in education: Testing the tests.* Rowley, MA: Newbury House.

Weaver, W. W. and A. J. Kingston. 1963. A factor analysis of the cloze procedure and other measures of reading and language ability. *The Journal of Communication 13*, 4:252-261.

APPENDIX A

Cloze Test

The discovery of the cause of the disease, malaria, began in 1880 when a French physician described a malarial parasite obtained from the blood of one of his patients. Italian investigators later demonstrated —(1)— the disease could be transmitted —(2)— human to human by infected blood, —(3)— in the 1890's British and Italian scientists suggested —(4)— type of mosquito named anopheles as the transmitter of the disease. By 1900 —(5)— had been established that —(6)— theory was correct, by demonstrating that the —(7)— was acquired only from the bite of —(8)— infected anopheles mosquito, —(9)— that persons protected from —(10)— mosquito did not contract the disease even in regions where —(11)— was rife. The biological cycle of —(12)— parasite has now been described in sufficient detail to explain —(13)— quinine was an effective remedy, why —(14)— bite of an infected mosquito —(15)— not transmit the disease until several days —(16)—the mosquito had become infected, and why —(17)— kinds of mosquitoes than the —(18)— did not transmit the disease.

—(19)— are four subtypes of the malarial parasite, —(20)— is known as plasmodium, —(21)— of which has its own characteristic pattern of biological changes as —(22)— passes from its —(23)— host, the mosquito, to its secondary —(24)—, man. The parasite undergoes sexual reproduction in the mosquito's stomach —(25)—. After a period of days, spores —(26)— seeds produced in the walls of the mosquito's stomach enter —(27)— salivary glands and are injected, along with saliva, into the bodies of future victims. —(28)— is only when this occurs that the —(29)— becomes capable of spreading the —(30)—.

Answer Key

STX	1.	that	STX	16. after
STX	2.	from	STX	17. other
COH	3.	and	COH	18. anopheles
STX	4.	*the*, a, one	STX	19. There
STX	5.	it	STX	20. which
COH	6.	the, *this*, their	COH	21. each
COH	7.	*disease*, parasite	COH	22. it
STX	8.	an	STG	23. primary
COH	9.	*and*, so	STG	24. host
COH	10.	the, *this*, that	STG	25. *walls*, lining, wall, cavity, tract
COH	11.	it, *malaria*, infection	COH	*26. or
COH	12.	the	COH	27. the *its*
STG	13.	why	STX	28. it
COH	14.	the	COH	29. mosquito, *insect*, anopheles
STX	15.	*did*, could	COH	30. *disease*, parasite

COH = Cohesive item
STX = Syntactic item
STG = Strategic item

(Underlined words occur in original passage.)

* Item 26 was not included in scoring, because of its inaccuracy of content.

APPENDIX B

Correlations among Composite Scores, Fall 1980 Group
(* = sig at $p \leq .01$, df = 315)

	$\bar{X}$	S	STX 1	STX 2	STX 3	STX 4	COH 1	COH 2	COH 3	COH 4	COH 5	COH 6	COH 7	STG 1	STG 2
STX 1	.70	.37	1.000												
STX 2	.77	.31	.316*	1.000											
STX 3	.57	.31	.331*	.415*	1.000										
STX 4	.35	.36	.172*	.123	.111	1.000									
COH 1	.49	.38	.197*	.162*	.182*	.168*	1.000								
COH 2	.68	.33	.300*	.199*	.262*	.214*	.216*	1.000							
COH 3	.66	.38	.234*	.283*	.299*	.105	.113	.218*	1.000						
COH 4	.57	.38	.303*	.363*	.506*	.093	.214*	.183*	.214*	1.000					
COH 5	.53	.36	.267*	.233*	.222*	.141*	.144*	.383*	.177*	.165*	1.000				
COH 6	.56	.38	.334*	.313*	.331*	.141*	.178*	.293*	.319*	.374*	.232*	1.000			
COH 7	.41	.37	.212*	.227*	.312*	.056	.067	.099*	.238*	.215*	.127	.302*	1.000		
STG 1	.47	.38	.322*	.345*	.464*	.179*	.177*	.185*	.313*	.355*	.179*	.368*	.305*	1.000	
STG 2	.37	.37	.316*	.313*	.382*	.153*	.142*	.232*	.244*	.293*	.132*	.257*	.281*	.426	1.000

APPENDIX B (Continued)

Correlations among Composite Scores, Spring 1981 Group
(* = sig at $p \leq .01$, df = 101)

	$\bar{X}$	S	STX 1	STX 2	STX 3	STX 4	COH 1	COH 2	COH 3	COH 4	COH 5	COH 6	COH 7	STG 1	STG 2
STX 1	.77	.32	1.000												
STX 2	.80	.29	.146	1.000											
STX 3	.55	.30	.040	.247*	1.000										
STX 4	.32	.31	.135	.198	.119	1.000									
COH 1	.45	.37	.188	.173	.267*	.284*	1.000								
COH 2	.79	.28	.045	.053	.148	−.018	.129	1.000							
COH 3	.68	.36	.291*	.230*	.158	.143	.292*	.162	1.000						
COH 4	.54	.35	.213	.223	.361*	.114	.207	.215	−.018	1.000					
COH 5	.52	.33	−.065	.183	.166	.195	.216	.031	.046	.122	1.000				
COH 6	.50	.35	.154	.217	.313*	.069	.284*	.127	.155	.221	.173	1.000			
COH 7	.45	.35	.256*	.218	.259*	.242*	.190	−.083	.129	.240*	.011	.343*	1.000		
STG 1	.44	.34	.229*	.192	.182	.244*	.303*	−.117	.213	.311*	.148	.288*	.368*	1.000	
STG 2	.37	.36	.061	.145	.234*	.118	.225	.040	.091	.378*	−.036	.236*	.285*	.334*	1.000