

SURFACE AND DEEP STRUCTURE OF LANGUAGE TESTS¹

Kanchana Prapphal

In order to measure to what extent a language learner knows the language, tests must be theoretically sound, reliable, and valid. Oller defines knowing a language as

...having internalized its grammar (where grammar is broadly defined to include pragmatic factors). Surely it does not mean simply knowing something about the language or having analyzed something to do with the grammar of the language. Thus, to test knowledge of the language it is necessary to press into action the internalized (and largely subconscious) grammar of the learner (1980, pp. 489-490).

Among the tests which appear to measure the efficiency learner's internalized grammar are integrative tests such as cloze procedure, dictation, essay writing, oral interviews. These tests, unlike discrete-point tests, do not isolate the syntactic, semantic, and pragmatic components of the language. Cloze tests were originally used as measures of reading comprehension and were popularized by Taylor (1953). The term "cloze", a notion of Gestalt, refers to the human psychological tendency to fill in gaps in patterns. Such a test is usually constructed by deleting every fifth word from a passage of prose. It has been demonstrated that cloze tests are measures of discourse processing as well as lower-order language skills by Oller (1975), Chihara, Oller, Weaver, and Chavez-Oller (1977), and Brown (1980).

Cloze tests have also been reported to be reliable indicators of non-native language proficiency by many researchers. Darnell (1970) found a reliability of .86 for a cloze test and a correlation of .83 between that test and the Test of English as a Foreign Language (TOEFL). Oller and Conrad (1971) found that the multiple correlation between all sections of the UCLA ESL Placement Examination (Form 2C) and the cloze test was .88. Perkins (1980) reported a correlation of .82 between a

¹ The author gratefully acknowledges Prof. John W. Oller, Jr. for his comments on an earlier version of this paper.

cloze test and the Michigan Test of English Language Proficiency (Form G) and a correlation of .80 between his cloze test and the Michigan Test (Form H). Farhady (1980) claimed that internal consistency estimates of reliability (KR-20, 21, and Cronbach's alpha) are inappropriate and tend to overestimate the reliability of cloze tests and other discourse-based tests. This, he claims, is due to the demonstrated dependency across items in such tests. However, as Brown (1980) pointed out, Farhady's claims do not show unreliability in cloze tests, but merely a possible inapplicability of certain methods of estimating reliability in such tests.

The validity of various language tests aimed at specific components have been questioned by Oller (1978, 1979). He points out that language tests called by different names may in fact measure the same thing to a great extent. Their considerable common variance may be due to a general factor of "intelligence" as postulated by Spearman (1904). Whatever nonoverlapping variance remains, may possibly be related to error variance and to certain surface aspects of language processing. Evidence for this possibility has been given in various studies: Stump (1978), Streiff (1978), and Flahive (1980). Nevertheless, Bachman and Palmer (1980) used confirmatory factoring methods to illustrate that there are actually additional factors beyond the general factor. Thus, we must allow for multiple-component models of language proficiency. This study investigates the relationship of different English language tests taken by Thai students and their high school grade-point average (GPA).

Method

Subjects

Five hundred and twenty-eight subjects from 13 departments at Chulalongkorn University in Bangkok, Thailand participated in the study. There were 63 Arts students, 64 Engineering, 105 Commerce and Accounting, 140 Science, 9 Political Science, 28 Law, 30 Education, 30 Medicine, 5 Veterinary Medicine, 2 Communication Arts, 17 Economics, 18 Dentistry, and 17 Pharmacy students. These students had studied English as a foreign language in a formal environment for 10 years.

Materials

The English Language Cloze Tests. The tests comprised two 250-word passages with every fifth word deleted. One passage dealt with social science and the other with physical science. There were 100 deleted items. The social science passage was rated by the Flesch readability formula as "fairly easy" and the physical science passage was rated as "easy". The instructions were given in Thai. Examples were given to make sure that the students would write only one word in each blank. The tests were scored by the contextually-appropriate method as suggested by Oller (1979) in order to obtain the maximum amount of meaningful variance. The reliability of the cloze tests was .599 and .762 for the two tests taken separately by the split-half method, and .765 for the two tests combined.

The Standardized Test : the Michigan Test of English Language Proficiency, Form K. The test was in multiple-choice format consisting of three subparts : grammatical usage, vocabulary and reading comprehension. There were 40 items in the first two parts and 20 in the last. The reliability of the test was .914 (KR-21).

The Chulalongkorn University Language Institute (CULI) Achievement Tests. The tests consisted of reading and listening based on criterion-referenced skills. The maximum scores for the two tests were 70 and 50, respectively. They were in multiple-choice format. Scores from the CULI were also obtained.

Results and Discussion

Number of cases, total scores, mean scores, and standard deviations of the CULI Achievement Tests, Michigan subtests, the Cloze Tests, and High School GPA are given in Table 1.

Table 1
 Number of Cases, Maximum Score, Mean Scores and Standard
 Deviations on the CULI Achievement Tests, Michigan Subtests, Cloze Tests, and GPA

Variable	N of Cases	Maximum Score	\bar{X}	SD
CULI Achievement Tests				
1. Reading	522	70	57.653	7.273
2. Listening	523	50	35.436	7.542
Michigan Test				
3. Structure	528	40	21.771	5.224
4. Vocabulary	528	40	16.081	4.519
5. Comprehension	528	20	9.233	3.140
6. Michigan Total (3+4+5)	528	100	47.085	10.676
Cloze Tests				
7. Cloze 1 (social science)	528	50	34.227	5.771
8. Cloze 2 (physical science)	528	50	27.953	7.674
9. Cloze Total (7+8)	528	100	62.197	12.646
10. High School GPA	463	4.0	2.949	.464

According to data given in the Michigan Test manual (not displayed in Table 1) the mean score of the Thai subjects on the Michigan Test places them just above the 28th percentile of the reference group upon which the test was normed. The norm was based on scores of 2347 candidates from 6 native language backgrounds: Arabic, Chinese, Germanic, Indic, Japanese, and Spanish. Thus the Thai subjects in this study were well below the midpoint of the norming population in English proficiency.

The social science cloze passage appears to be easier than the science passage. The mean score for the former is 34.227 and the latter is 27.953 (out of 50 in each case). This is contradictory to the prediction of the Flesch formula which rated the social science passage as “fairly easy” and the science as “easy”. This calls into question the validity of ratings based on a readability formula which refers only to surface structures (e.g., such facts as the number of words and sentences in the text) while ignoring requirements pertaining to knowledge of the world. The results of the actual testing reveal that pragmatic aspects of the language, must also be taken into account when judging the difficulty level of the passages. The social science passage, “People Have the Same Needs”, was easier than the physical science passage, “Heating the Atmosphere”. This is probably because the former text appeals to more common knowledge than the latter.

The intercorrelations among the CULI Achievement, Michigan, Cloze Tests, and High School GPA are presented in Table 2.

Table 2

Intercorrelations among the CULJ Achievement Tests,
Michigan Subtests, Cloze Tests and GPA

Variable	1	2	3	4	5	6	7	8	9	10
CULJ Achievement Tests										
1. Reading (522)	1.000.	.546	.574	.473	.427	.607	.607	.655	.677	.447
2. Listening (523)		1.000	.555	.463	.496	.614	.558	.651	.653	.380
Michigan Test										
3. Structure (528)			1.000	.578	.467	.871	.568	.645	.652	.472
4. Vocabulary (528)				1.000	.486	.849	.428	.524	.513	.368
5. Comprehension (528)					1.000	.729	.391	.490	.477	.312
6. Michigan Total : 3+4+5 (528)						1.000	.574	.682	.676	.480
Cloze Tests										
7. Cloze 1, social science (528)							1.000	.764	.919	.449
8. Cloze 2, physical science (528)								1.000	.956	.472
9. Cloze Total : 7+8 (528)									1.000	.490
10. High School GPA (463)										1.000

All are significant at .001 (2-tailed test). () = N

The intercorrelations between the CULI Achievement Tests, the subparts of the Michigan Test, the Cloze Tests, and High School GPA are all significant at the .001 probability level (2-tailed test). In terms of common variance, the Cloze Total accounts for 46% of the variance in the Michigan Total. The Cloze Total has a common variance of 46% with the CULI Reading Test and 43% with the Listening Test. Similarly, the Michigan Total and the CULI Reading Test share a maximum common variance of about 37% and 38% with the Listening Test. The common variances between High School GPA with the Reading, Listening, Michigan Total, and Cloze Total are 20%, 14%, 23%, and 24% respectively. The common variance among the tests and past academic performance indicates that the tests and GPA may be measuring somewhat the same thing (s) at a deep level. This appears to be true when looking at the underlying structure of the tests and GPA as given in Table 3.

Table 3

Factor Matrix Using Principal Factor with Iterations of High School GPA,
the CULI Achievement Tests, Michigan Subtests and Cloze Tests

Variable	General Factor	Communality
1. High School GPA	.545	.298
CULI Achievement Tests		
2. Reading	.762	.580
3. Listening	.746	.556
Michigan Test		
4. Structure	.786	.619
5. Vocabulary	.659	.434
6. Comprehension	.601	.362
Cloze Tests		
7. Cloze 1 (social science)	.775	.601
8. Cloze 2 (physical science)	.878	.771
		<u>.771</u>
	Eigenvalue =	4.221

The varimax rotated solution with iterations presented in Table 3 shows all of the variables loading fairly strongly on one factor. This factor may be very similar to the general factor of intelligence as proposed by Spearman--a deep language factor as suggested by Oller or possibly some strong group factor. Further evidence for underlying structure of the tests and cognitive ability is obtained from a canonical correlation analysis given in Table 4.

Table 4

Canonical Correlation of the CULI
Achievement Tests, Michigan Subtests, Cloze Tests, and High School GPA

Number	Eigenvalue	Canonical Correlation	Wilk's Lambda	Chi-square	D.F.	Significance
1	.656	.810	.337	492.344	12	0.000
2	.020	.141	.980	9.074	5	0.106
Coefficients for canonical variables of the first set						
Canvar 1						
Structure		.302				
Vocabulary		.084				
Comprehension		.102				
Reading		.379				
Listening		.308				
High School GPA		.088				
Coefficients for canonical variables of the second set						
Canvar 1						
Cloze 1 (social science)		.300				
Cloze 2 (physical science)		.752				

This analysis reveals that there is one significant relationship among the tests and GPA. The first set consists of the tests in a multiple-choice format, the CULI Achievement Tests, the Michigan subtests, and GPA. The second set is the open-ended format (Cloze Tests). The two sets are related to each other in one significant dimension sharing a total of 66% of their total variance. This lends support to the hypothesis that language tests and academic performance may be intricately related at their deepest levels.

Conclusion

Examination of the language tests used in this study indicates that the distinct categories of the subtests are perhaps meaningful only at the surface level but less so at the underlying level. The communality among the tests may represent global knowledge of language as distinct from different surface skills represented in the various subtests. The underlying global trait may be an ability to manipulate propositional forms, an ability necessary to all of the measures. It is also notable that there is clearly a substantial relation between the tests and overall GPA. It seems axiomatic that the deep cognitive machinery involved in a general academic performance may also play a major role in the learning of English.

References

- Bachman, L.F. and Palmer, A.S. Trait and method factors in language proficiency test scores. Presented at the Language Testing Conference of the LSA/TESOL Summer Institutes, Albuquerque, June 1980. Also in J.W. Oller, Jr. (Ed.), *Issues in language testing research*. Rowley, Mass. : Newbury House, in press.
- Brown, J.D. A closer look at cloze. Paper presented at the Language Testing Conference of the LSA/TESOL Summer Institutes, Albuquerque, June 1980. Appears in shortened form in J.W. Oller, Jr. (Ed.), *Issues in language testing research*. Rowley, Mass. : Newbury House, in press.
- Chihara, T., Oller, J.W. Jr., Weaver, K., & Chavez-Oller, M.A. Are cloze items sensitive to constraints across sentences? *Language Learning*, 1977, 27, 63-73.
- Darnell, D.K. The development of an English language proficiency test of foreign students using the clozentropy procedure. *Speech Monographs*, 1970, 37, 36-46.

- Farhady, H. New directions for ESL proficiency testing. Presented at the Language Testing Conference of the LSA/TESOL Summer Institutes, Albuquerque, June 1980. Also in J.W. Oller, Jr. (Ed.), *Issues in language testing research*. Rowley, Mass. : Newbury House, in press.
- Flahive, D.E. Separating the g factor from reading comprehension. In J.W. Oller, Jr., & K. Perkins (Eds.), *Research in language testing*. Rowley, Mass. : Newbury House, 1980.
- Flesch, R. A new readability yardstick. *Journal of Applied Psychology*, 1948, 32, 221-233.
- Oller, J.W. Jr. Cloze, discourse, and approximations to English. In M.K. Burt & H.C. Dulay (Eds.), *New directions in second language teaching and bilingual education : on TESOL' 75*. Washington, D.C. : TESOL, 1975.
- Oller, J.W. Jr. How important is language proficiency to IQ and other educational tests? In J.W. Oller, Jr. & K. Perkins (Eds.), *Language in education : testing the tests*. Rowley, Mass. : Newbury House, 1978.
- Oller, J.W. Jr. *Language tests at school*. London : Longman, 1979.
- Oller, J.W. Jr. Language testing today. In K. Croft (Ed.), *Readings on English as a second language* (2nd ed.). Cambridge, Mass : Winthrop Publishers, Inc., 1980.
- Oller, J.W. Jr., & Conrad, C. The cloze procedure and ESL proficiency. *Language Learning*, 1971, 21, 183-196.
- Perkins, K. A statistical analysis of different instruments to measure short-term gains in an L2 immersion program. Paper read at the 1980 MID TESOL Meeting at the University of Missouri at Kansas City, 1980.
- Spearman, C.E. 'General intelligence' objectively determined and measured. *American Journal of Psychology*, 1904, 15, 201-293.
- Streiff, V. Relationships among oral and written cloze scores and achievement test scores in a bilingual setting. In J.W. Oller, Jr. & K. Perkins (Eds.). *Language in education : testing the tests*. Rowley, Mass. : Newbury House, 1978.
- Stump, T.A. Cloze and dictation tasks as predictors of intelligence and achievement scores. In J.W. Oller, Jr. & K. Perkins (Eds.), *Language in education : testing the tests*. Rowley, Mass. : Newbury House, 1978.
- Taylor W.L. Cloze procedure : a new tool for measuring readability. *Journalism Quarterly*, 1953, 30, 415-433.