DOCUMENT RESUME

ED 365 152 FL 021 770

AUTHOR	Brown, James Dean; And Others
TITLE	Southeast Asian Languages Proficiency Examinations.
PUB DATE	91
NOTE	21p.; In: Sarinee, Anivan, Ed. Current Developments in Language Testing. Anthology Series 25. Paper presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (April 9-12, 1990); see FL 021 757.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cambodian; Cloze Procedure; Comparative Analysis; Dictation; Indonesian; Interviews; *Language Proficiency; *Language Tests; Listening Comprehension; *Second Languages; Tagalog; *Test Construction; Test Format; Test Reliability; Test Validity; Thai; *Uncommonly Taught Languages; Vietnamese
IDENTIFIERS	ACTFL Proficiency Guidelines; Southeast Asian Summer Studies Institute

ABSTRACT

The design, administration, revision, and validation of the Southeast Asian Summer Studies Institute proficiency examinations are reported. The examinations were created as parallel language proficiency tests in each of five languages: Indonesian, Khmer, Tagalog, Thai, and Vietnamese. Four tests were developed in each language: multiple-choice listening comprehension, interview, dictation, and cloze. The interview and listening tests were each designed to assess all of the levels of language ability in the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines. The study reported here (involving 218 students) explored the score distributions for each test on the proficiency batteries for each language, as well as differences between distributions for the pilot and revised versions. Relative reliability estimates for pilot and revised versions and the relationships of tests across languages were also compared. Based on the analyses it is concluded that the tests in each of the five examinations are reasonably well-centered and reliable, and distributions are adequate. (MSE)

SOUTHEAST ASIAN LANGUAGES PROFICIENCY EXAMINATIONS

James Dean Brown
H. Gary Cook
Charles Lockhart
Teresita Ramos

ABSTRACT

This paper reports on the design, administration, revision and validation of the Southeast Asian Summer Studies Institute (SEASSI) Proficiency Examinations. The goal was to develop parallel language proficiency examinations in each of five languages taught in the SEASSI: Indonesian, Khmer, Tagalog, Thai and Vietnamese. Four tests were developed for each of these languages: multiple-choice listening, interview, dictation and cloze test. To maximize the relationships among these examinations and the associated curricula, the interview and listening tests were each designed to assess all of the levels of language ability which are described in the *ACTFL Proficiency Guidelines* from "novice" to "advanced-plus."

This study (N = 218) explored the score distributions for each test on the proficiency batteries for each language, as well as differences between the distributions for the pilot (1989) and revised (1989) versions. The relative reliability estimates of the pilot and revised versions were also compared as were the various relationships among tests across languages.

The results are discussed in terms of the degree to which the scores on the strategies here are generalizable to test development projects for other Southeast Asian languages.

Each year since 1984, a Southeast Asian Summer Studies Institute (SEASSI) has been held on some university campus in the United States. As the name implies, the purpose of SEASSI is to provide instruction in the "lesser taught" languages from Southeast Asia. In 1988, SEASSI came to the university of Hawaii at Manoa for two consecutive summers. Since we found ourselves with several language testing specialists, a strong Indo-Pacific Language department, and two consecutive years to work, we were in a unique position to develop overall proficiency tests for a number of the languages taught in SEASSI -- tests that could then be passed on to future SEASSIs.

The central purpose of this pap.. is to describe the design, production, administration, piloting, revision and validation of these Southeast Asian Summer Studies Institute Proficiency Examinations (SEASSI). From the outset, the goal of this project was to develop overall language proficiency examinations in each of five languages taught in the SEASSI: Indonesia, Khmer, Tagalog, Thai and Vietnamese. The ultimate objectives of these tests was to assess the grammatical and communicative ability of students studying these languages in order to gauge their overall proficiency in the languages. It was decided early that the tests should be designed to measure all of the levels of language ability which are described in the ***ACTFL Proficiency Guidelines*** from "novice" to "advanced-plus" for speaking and listening (see Appendix A from ACTFL 1986, Liskin-Gasparro 1982, and/or ILR 1982). Though the ACTFL guidelines are somewhat controversial (eg. see Savignon 1985; Bachman and Savignon 1986), they provided a relatively simple paradigm within which we could develop and describe these tests in terms familiar to all of the teachers involved in the project, as well as to any language teachers who might be required to use the tests in the future.

The central research questions investigated in this were as follows

(1) How are the scores distributed for each test of the proficiency battery for each language, and how do the distributions differ between the pilot (1989) and revised (1989) versions?

(2) To what degree are the tests reliable? How does the reliability differ between the pilot and revised versions?

(3) To what degree are the tests intercorrelated? How do these correlation coefficients differ between the pilot and revised versions?

(4) To what degree are the tests parallel across languages?

(5) To what degree are the tests valid for purposes of testing overall proficiency in these languages?

(6) To what degree are the strategies described here generalizable to test development projects for other languages?

METHOD

A test development project like this has many facets. In order to facilitate the description and explanation of the project, this **METHOD** section will be organized into a description of the subject used for norming the tests, a section on the materials involved in the testing, an explanation of the procedures of the statistical procedures used to analyze, improve and reanalyze the tests.

Subject

A total of 228 students were involved in this project: 101 in the pilot stage of this project and 117 in the validation stage.

The 101 students involved in the pilot stage were all students in the SEASSI program during the summer of 1989 at the University of Hawaii at Manoa. They were enrolled in the first year (45.5%), second year (32.7%) and third year (21.8%) language courses in Indonesian (n = 26), Khmer (n = 21), Tagalog (n = 14) Thai (n = 17) and Vietnamese (n = 23). There were 48 females (47.5%) and 53 Males (52.5%). The vast majority of these students were native speakers of English (80.7%), though there were speakers of other languages who participated (19.3%).

The 117 students involved in the validation stage of this test development project were all students in the SEASSI program during summer 1989. They were enrolled in the first year (48.7%), second year (41.0%) and third year (10.3%) language courses in Indonesian (n = 54), Khmer (n = 18), Tagalog (n = 10) Thai (n = 23) and Vietnamese (n = 12). There were 57 females (48.7%) and 60 males (51.3%).

In general, all of the groups in this study were intact classes. To some degree, the participation of the students depended on the cooperation of their teachers. Since that cooperation was not universal, the samples in this project can only be viewed as typical of volunteer groups drawn from a summer intensive language study situation like that in SEASSI.

Materials

There were two test batteries employed in this project. The test of focus was the SEASSIPE. However, the ***Modern Language Aptitude Test*** (MLAT), developed by Carroll and Sapon (1959), was also administered. Each will be described in turn.

Description of the SEASSIPE. The SEASSIPE battery for each language presently consisted of four tests : multiple-choice listening, oral interview procedure, dictation and cloze test. In order to make the tests as comparable as possible across the five languages, they were all developed first in an English prototype version. The English version was then translated into the target language with an emphasis on truly translating the material into that language such that the result would be natural Indonesian, Khmer, Tagalog, Thai or Vietnamese. The multiple-choice ***listening*** test presented the students with aural statements or questions in the target language, and they were then asked what they would say (given four responses to choose from). The pilot versions of the test all contained 36 items, which were developed in 1988 on the basis of the ACTFL guidelines for listening (see **APPENDIX A**). The tests were then administered in the 1988 SEASSI. During 1989, the items were revised using distractor efficiency analysis, and six items were eliminated on the basis of overall item statistics. Thus the revised versions of the listening test all contained a total of 30 items.

The ***oral interview*** procedure was designed such that the interviewer would ask students questions at various levels of difficulty in the target language (based on the ACTFL speaking and listening guidelines in **APPENDIX A**). The students were required to respond in the target language. In the pilot version of the test, the responses of the students were rated on a 0-108 scale. On each of 36 questions, this scale had 0 to 3 points (one each for three categories: accuracy, fluency, and meaning). On the revised version of the interview, 12 questions were eliminated. Hence on the revised version, the students were rated on a 0-72 scale including one point each for accuracy, fluency and meaning based on a total of 24 interview questions.

The ***dictation*** consisted of an eighty word passage in the target language. The original English prototype was of approximately 7th grade reading level (using the ***Fry*** 1976 scale). The passage was read three times (once at normal rate of speech, then again with pauses at the end of logical phrases, and finally, again at normal rate). Each word that was morphologically correct was scored as a right answer. Because these dictations appeared to be working reasonably well, only very minor changes were made between the pilot and revised versions of this test.

The ***cloze*** test was based on an English prototype of 450 words at about the 7th grade reading level (again using the Fry 1976 scale). The cloze passage was created in the target language by translating the English passage and deleting every 13th word for a total of 30 blanks. The pilot and revised versions of this test each had the same number of items. However, blanks that proved ineffective statistically or linguistically in the pilot versions were changed to more promising positions in the revised tests (see Brown 1988b for more on cloze test improvement strategies).

As mentioned above, these four tests were developed for each of five languages taught in the SEASSI. To the degree that it was possible, they were made parallel across languages. The goal was that scores should be comparable across languages so that, for instance, a score of 50 on the interview procedure for Tagalog would be approximately the same as a score of 50 on the Thai test. To investigate the degree to which the tests were approximately equivalent across languages, the ***Modern Language Aptitude Test*** was also administered at the beginning of the instruction so that the results could be used to control for initial differences in language aptitude among the language groups.

All of the results of the SEASSI Proficiency Educations were considered experimental. Hence the results of the pilot project were used primarily to improve the tests and administration procedures in a revised version of each test. The scores were reported to the teachers to help in instructing and grading the students. However, the teachers were not required, in any way, to use the results, and the results were NOT used to judge the effectiveness of instruction. Teachers' input was solicited and used at all points in the test development process.

Description of the MLAT. The short version of the MLAT was also administered in this study. Only the last three of the five tests were administered as prescribed for the short version by the original authors. These three tests are entitled ***spelling clues, words in sentences and paired associates***.

The MLAT was included to control for differences in language learning aptitude across the five language groups and thereby help in investigating the equivalency of the tests across languages. The MLAT is a well-known language aptitude test. It was designed to predict performance in foreign language classroom. In this study, the results were kept confidential and did not affect the students' grades in any way. The scores and national percentile ranking were reported individually to the students with the caution that such scores represent only one type of information about their aptitude for learning foreign languages. It was made clear that the MLAT does not measure achievement in a specific language. The group scores, coded under anonymous student numbers, were only used to make general observations and to calculate some of the statistical analyses reported below.

Procedures

The overall plan for this project proceeded on schedule in four main stages and a number of smaller steps.

Stage one: Design. The tests were designed during June 1988 at the University of Hawaii at Manoa by J D Brown, Charles Lockhart and Teresita Ramos with the cooperation of teachers of the five languages involved (both in the Indo-Pacific Languages department and in SEASSI). J D Brown and C Lockhart were responsible for producing a prototypes into each of the five languages. J D Brown took primary responsibility for overall test design, administration and analysis.

Stage two: Production. The actual production of the tapes, booklets, answer sheets, scoring protocols and proctor instructions took place during the last week of July 1988 and the tests were actually administered in SEASSI classes on August 5, 1988. This stage was the responsibility of T. Ramos with the help of C. Lockhart.

Stage three: Validation. The on-going validation process involved the collection and organization of the August 5th data, as well as teacher ratings of the students' proficiency on the interview. Item analysis, descriptive statistics, correlational analysis and feedback from the teachers and students were all used to revise the four tests with the goal of improving them in terms of central tendency,dispersion, reliability and validity. The actual revisions and production of new versions of the tests took place during the spring and summer of 1989. This stage was primarily the responsibility of J D Brown with the help and cooperation of H Gary Cook, T Ramoa and the SEASSI teachers.

Stage four: Final Product. Revised versions of these tests were administered again in the 1989 SEASSI. This was primarily the job of H G Cook. A test manual was also produced (Brown, Cook, LocKhart and Ramos, unpublished ms). Based on the students' SEASSI performances and MLAT scores from both the 1988 and 1989 SEASSI, the manual provides directions for administering the tests, as well as discussion of the test development and norming procedures. The discussion focuses on the value of these new measures as indirect tests of ACTFL proficiency levels. The manual was developed following the standards set by AERA, APA and NCME in *Standards for Educational and Psychological Testing* (see APA 1985). The production of all tests, answer keys, audio tapes, answer sheets, manuals and reports was the primary responsibility of J D Brown.

Analyses

The analyses for this study were conducted using the *QuattroPro* spreadsheet program (Borland 1989), as well as the *ABSTAT* (Bell-Anderson 1989), and *SYSTAT* (Wilkinson, 1988) statistical program. These analyses fall into four categories: descriptive statistics, reliability statistics, correlational analyses, and analysis of covariance.

Because of the number of tests involved when we analyzed four tests each in two versions (1988 pilot version and 1989 revised version) for each of five languages (4 x 5 x 2 = 40), the *descriptive statistics* reported here are limited to the number of items, the number of subjects, the mean and the standard deviation. Similarity, ***reliability statistics*** have been limited to the Cronbach alpha coefficient (sec Cronbach 1970) and the Kuder and Richardson (1973) formula 21 (K-R21). All ***correlation coefficients*** reported here are Pearson product-moment coefficients. Finally, ***analysis of covariance*** (ANCOVA) and multivariate analyses were used to determine the degree while controlling for differences in initial language aptitude (as measured by the MLAT). the alpha significance level for all statistical decisions was set at .05.

RESULTS

Summary descriptive statistics are presented in Table 1 for the pilot and revised versions of the four tests for each of the five languages. The languages are listed across the top of the table with the mean and standard deviation for each given directly below the language headings. The mean provides an indication of the overall central tendency, or typical behavior of a group, and the standard deviation gives an estimate of the average distance of students from the mean (see Brown 1988a for more on such statistics). The versions (ie. the pilot versions administered in summer of 1988 or the revised versions administered in summer of 1989) and tests (Listening, Oral, Interview, Dictation and Cloze Test) are labeled down the left side of the table along with the number of items (k) in parentheses.

TABLE 1: SEASSIPE DESCRIPTIVE STATISTICS

VERSION TEST	INDONESIAN		KHMER		TAGALOG		THAI		VIETNAMESE	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
PILOT 1988										
Listening (k = 32)	22.55	4.14	17.95	5.12	20.00	6.75	13.76	5.61	17.87	5.25
Oral Intv (k = 178)	[illegible]	20.50	77.84	23.56	63.92	20.63	83.23	9.02	74.08	18.15
Dictation (k = 80)	37.67	8.12	35.60	5.59	33.83	10.96	17.38	9.86	29.67	12.35
Cloze Tst (k = 30)	13.52	4.97	9.00	4.06	12.23	5.75	16.50	3.07	15.64	6.57
REVISED 1989										
Listening (k = 30)	20.38	3.31	17.72	5.20	17.70	3.77	11.77	4.57	20.75	3.96
Oral Intv (k = 72)	64.59	6.66	61.89	9.55	56.50	10.32	50.22	13.92	57.33	6.31
Dictation (k = 80)	49.96	9.70	33.20	14.77	49.90	22.83	36.86	9.10	68.58	7.03
Cloze Tst (k = 30)	20.09	4.12	17.90	4.15	17.90	6.33	13.50	4.34	14.50	5.30

Notice that, for each test, there is considerable variation across versions and languages not only in the magnitude of the means but also among the standard deviations. It seems probable that the disparities across versions (1988 and 1989) are largely due to the revision processes, but they may in part be caused by differences in the numbers of students at each level of study or by other differences among the samples used during the two summers.

Table 2 presents the reliabilities for each test based on the scores produced by the groups of students studying each of the languages. A reliability coefficient estimates the degree to which a test is consistent in what it measures. Such coefficient can range from 0.00 (wholly unreliable, or inconsistent) to 1.00 (completely reliable, or 100 percent consistent), and can take on all of the values in between, as well.

Notice that, once again, the languages are shown across the top of the table with two types of reliability, alpha and k-R21, labeled just under each language heading. You will also find that the versions (1988 or 1989) and tests are again labeled down the left side of the table.

TABLE 2: SEASSI TEST RELIABILITY FOR EACH LANGUAGE

VERSION	INDONESIAN		KHMER		TAGALOG		THAI		VIETNAMESE	
Test	ALPHA	K-R21	ALPHA	K-R21	ALPHA	K-R21	ALPHA	K-R21	ALPHA	K-R21
PILOT 1988										
Listening	.79	.63	.78	.72	.91	.86	.85	.78	.82	[illegible]
Oral Intv	.97	.96	.99	.97	[illegible]	[illegible]	.76	.76	.98	[illegible]
Dictation	**	.80	**	.89	**	.91	**	.50	**	.75
Cloze Tst	*	.72	*	.64	*	.81	*	.22	*	.86
REVISED 1989										
Listening	[illegible]	.42	.92	.76	.69	.51	.81	.68	.76	[illegible]
Oral Intv	.91	.86	.94	.92	.95	.90	.97	.93	.78	.72
Dictation	**	.81	**	.92	**	.98	**	.78	**	.91
Cloze Tst	.77	.63	.97	.60	.96	.85	.99	.63	.84	.76

* Not calculated.
** Not applicable.

As mentioned above, the reliability estimates reported in Table 2 are based on Cronbach alpha and on the K-R21. Cronbach alpha is an algebraic identity with the more familiar K-R20 for any test which is dichotomously scored (eg. the listening and cloze tests in this study). However, for any test which has a weighed scoring system (like the Interview tests in this study), another version of alpha must be applied -- in this case, one based on the odd-even variances (see Cronbach 1970)

TABLE 3: SEASSIPE TEST INTERCORRELATIONS FOR EACH LANGUAGE

VERSION	INDONESIAN			KHMER			TAGALOG			THAI			VIETNAMESE		
Tests	L	O	D	L	O	D	L	O	D	L	O	D	L	O	D
PILOT 1988															
Oral Intv	.54*			.64			.83*			.26			.65*		
Dictation	.60*	.70*		.60*	.46*		.92*	.83*		-.12	.42		.54*	.62*	
Cloze Tst	.57*	.79*	.82*	.80	.57*	.14*	.63*	.55*	.67*	-.24	.20	.94*	.73*	.75*	.67*
REVISED 1989															
Oral Intv	.51*			.70*			.72*			.41			.59*		
Dictation	.38*	.58*		.63*	.64*		.56	.83*		-.69	-.71		.82*	.77*	
Cloze Tst	.30	.24	.58*	.76*	.74*	.01*	.69*	.92*	.79*	.56	.50	-.19	.73*	.77*	.72*

* p < .05
*L = LISTENING; O = ORAL INTV; D = CLOZE TST

Intercorrelations among the SEASSIPE tests on both versions were calculated using the Pearson product-moment correlation coefficient for each language separately (see Table 3). A correlation coefficient gives an estimate of the degree to which two sets of numbers are related. A coefficient of 0.00 indicates that the numbers are totally unrelated. A coefficient of +1.00 indicates that they are completely related (mostly in terms of being ordered in the same way). A coefficient of -1.00 indicates that they are strongly related, but in opposite directions, ie. as one set of numbers becomes larger, the other set grows smaller. Naturally, coefficients can vary throughout this range from -1.00 to 0.00 to 1.00.

Notice that the languages are labeled across the top with Listening (L), Oral Interview (O) and Dictation (D) also indicated for each language. The versions (1988 or 1989) and tests (Oral Interview, Dictation and Cloze Test) are also indicated down the left side. To read the table, remember that each correlation coefficient is found at the intersection of the two variables that were being examined. This means, for instance, that the .54 in the upper-left corner indicates the degree of relationship between the scores on the Oral Interview and Listening tests in Indonesian in 1988 pilot version.

Following some of the correlation coefficients in Table 3, there is an asterisk, which refers down below the table to p < .05. This simply means that these correlation coefficients are statistically significant at the .05 level. In other words, there is only a five percent probability that the correlation coefficients with asterisks occurred by chance alone. Put yet another way, there is a 95 percent probability that the coefficients with asterisks occurred for other than chance reasons. Those coefficients without asterisks can be interpreted as being zero.

Recall that, in Table 1, there was considerable variation in the magnitude of the means and standard deviations across languages and versions. Table 4 shows the results of an analysis of covariance procedure which used language (Indonesian, Khmer, Tagalog, Thai and Vietnamese) as a categorical variable and MLAT language aptitude scores as a covariate to determine whether there were significant differences across languages for the mean test scores (Listening, Interview, Dictation and Cloze treated as repeated measures).

TABLE 4: ANALYSIS OF COVARIANCE ACROSS REPEATED MEASURES (TESTS)

SOURCE	SS	df	MS	F
BETWEEN SUBJECTS				
LANGUAGE	3197.197	4	799.299	7.248*
MLAT (COVARIATE)	256.014	1	256.014	2.322
SUBJECTS WITHIN GROUPS	6285.642	57	110.274	
WITHIN SUBJECTS				
LANGUAGE	7156.650	12	596.387	18.196*
MLAT (COVARIATE)	80.513	3	26.838	0.819
SUBJECTS WITHIN GROUPS	5604.643	171	32.776	

*p < .05

In Table 4, it is important to realize that the asterisks next to the F ratios indicate that there is some significant difference among the means for different languages across the four tests. This means in effect that at least one of the differences in means shown in table 1 is due to other than chance factors (with 95 percent certainly). Of course, many more of the differences may also be significant, but there is no way of knowing which they are from this overall analysis. It should suffice to recognize that a significant difference exists somewhere across languages. The lack of asterisks after the F ratios for the MLAT indicate that there was no significant difference detected language aptitude (as measured by MLAT) among the groups of students taking the five languages.

Since analysis of covariance is a fairly controversial procedure, two additional steps were taken:

(1) First, the assumption of homogeneity of slopes was carefully checked by calculating and examining the interaction terms before performing the actual analysis of covariance. The interactions were not found to be significant.

(2) Second, multivariate analyses (including, Wilks' lambda, Pillai trace, and Hotelling-Lawley trace) were also calculated. Since they led to exactly the same conclusions as the univariate statistics shown in Table 4, they are not reported here.

Thus the assumptions were found to be met for the univariate analysis of covariance procedures in a repeated measures design, and the results were further confirmed using multivariate procedures. It is therefore with a fair amount of confidence that these results are reported here.

TABLE 5: DIFFERENTIAL PERFORMANCE BY LEVELS ON EACH TEST

TEST	LEVEL	MEAN	STD	N
Listening	1st year	15.7347	5.2392	49
	2nd year	19.6383	4.4007	47
	3rd year	20.9167	4.5218	12
Oral Intv	1st year	50.6538	14.9022	26
	2nd year	47.1915	15.9519	47
	3rd year	57.5000	12.2734	12
Dictation	1st year	16.4063	5.3573	32
	2nd year	18.2500	4.2602	48
	3rd year	23.9167	3.4499	12
Cloze Tst	1st year	57.3393	12.1015	56
	2nd year	61.5000	8.8894	48
	3rd year	65.5833	6.8948	12

One other important result was found in this study: the tests do appear to reflect the differences in ability found between levels of language study. This is an important issue for overall proficiency tests like the SEASSIPE because they should be sensitive to the types of overall differences in language ability that would develop over time, or among individuals studying at different levels. While this differential level effect was found for each of the languages, it is summarized across languages in Table 5 (in the interests of economy of space). Notice that, with one exception, the means get higher on all of the tests as the level of the students goes up from first to second to third year. The one anomaly is between the first and second years on the oral interview.

DISCUSSION

The purpose of this section will be to interpret the results reported above with the goal of providing direct answers to the original research questions posed at the beginning of this study. Consequently, the research questions will be restated and used as headings to help organize the discussion.

(1) *How are the scores distributed for each test of the proficiency battery for each language, and how do the distributions differ between the pilot (1989) and revised (1989) versions?*

The results in Table 1 indicate that most of the current tests are reasonably well-centered and have scores that are fairly widely dispersed about the central tendency. Several notable exceptions seem to be the 1989 Oral Interviews for Indonesian and Khmer, both of which appear to be negatively skewed (providing classic examples of what is commonly called the ceiling effect -- see Brown 1988a for further explanation). It is difficult, if not impossible, to disentangle whether the differences found between the two versions of the test (1988 and 1989) are due to the revision processes in which many of the tests were shortened and improved, or to differences in the samples used during the two SEASSIs.

(2) *To what degree are the tests reliable? How does the reliability differ between the pilot and revised versions?*

Table 2 shows an array of reliability coefficients for the 1988 pilot version and 1989 revised tests that are all moderate to very high in magnitude. The lowest of these is for the 1989 Indonesian Listening test. It is low enough that the results for this test should only be used with extreme caution until it can be administered again to determine whether the low reliability is a result of bad test design or some aspect of the sample of students who took the test.

These reliability statistics indicate that most of the tests produce reasonably consistent results even when they are administered to the relatively homogeneous population of SEASSI students. The revision process appears to have generally, though not universally, improved test reliability either in terms of producing higher reliability indices or approximately equal estimates, but for shorter more efficient, versions. The listening tests for Indonesian and Tagalog are worrisome because the reliabilities are lower in the revised than in the pilot testing and because they are found among the 1989 results. However, it is important to remember that these are fairly short tests and that they are being administered to relatively restricted ranges of ability in the various languages involved. These are both important factors because, all things being equal, a short test will be less reliable than a long test, and a restricted range of talent will produce lower reliability estimates than a wide one (for further explanation and examples, see Ebel 1979; Brown 1984, 1988a).

Note also that the K-R21 statistic is generally lower than the alpha estimate. This is typical. K-R21 is a relatively easy to calculate reliability estimate, but it usually underestimates the actual reliability of the test (see, for instance, the 1989 Revised Khmer and Thai cloze tests reliabilities in Table 2).

(3) *To what degree are the tests intercorrelated? How do these correlation coefficients differ between the pilot and revised versions?*

In most cases, the correlation coefficients reported in Table 3 indicate a surprisingly high degree of relationship among the tests. The one systematic and glaring exception is the set of coefficients found for Thai. It is important to note that these correlation coefficients for Thai based on very small samples (due mostly to the fact that students at the lowest level were not taught to write in Thai), and that these correlation coefficients were not statistically significant at the $p < .05$ level. They must therefore be interpreted as correlation coefficients that probably occurred by chance alone, or simply as correlations of zero.

(4) *To what degree are the tests parallel across languages?*

The interpretation of these results is fairly straightforward. Apparently, there was no statistically significant difference in MLAT language aptitude scores among the groups studying the five languages. However, there was clearly a significant difference among the mean test scores across the five languages despite the efforts to control for initial differences in language aptitude (the MLAT covariate). A glance back at Table 1 will indicate the magnitude of such differences.

One possible cause for these differences is that the tests have changed during the process of development. Recall that all of these tests started out as the same English language prototype . It is apparent that, during the processes of translating and revising, the tests diverged in overall difficulty across languages. this is reflected in the mean differences found here. Another potential cause of the statistically significant differences reported in Tables 1, 4, and 5 is that there may have been considerable variations in th: samples used during the two summers.

(5) ***To what degree are the tests valid for purposes of testing overall proficiency in these languages?***

The intercorrelations among the tests for each language (see Table 3) indicate that moderate to strong systematic relationships exist among many of the tests in four of the five languages being tested in this project (the exception is Thai). However, this type of correlational analysis is far from sufficient for analysing the validity of these tests. If there were other well established tests of the skills being tested in these languages, it would be possible to administer those criterion tests along with the SEASSIPE tests and study the correlation coefficients between our relatively new tests and the well-established measures. Such information could then be used to build arguments for the criterion-related validity of some or all of these measures. Unfortunately, no such well-established criterion measures were available at the time of this project.

However, there are results in this study that do lend support to the construct validity of these tests. The fact that the tests generally reflect differences between levels of study (as shown in Table 3) provides evidence for the construct validity (the differential groups type) of these tests.

Nevertheless, much more evidence should be gathered on the validity of the various measures in this study. An intervention study of their construct validity could be set up by administering the tests before and after instruction to determine the degree to which they are sensitive to the language proficiency construct which is presumably being taught in the course. If, in future data, correlational analyses indicate patterns similar to those found here, factor analyses factor analysis might also be used profitably to explore the variance structures of those relationships.

The point is that there are indications in this study of the validity of the tests involved. However, in the study of validity, it is important to build arguments from a number of perspectives on an ongoing basis. Hence, in a sense, the study of validity is never fully complete as long as more evidence can be gathered and stronger arguments can be constructed.

(6) ***To what degree are the strategies described here generalizable to test development projects for other languages?***

From the outset, this project was designed to provide four different types of proficiency tests -- tests that would be comparable across five languages. The intention was to develop tests that would produce scores that were comparable across languages such that a score of 34 would be roughly comparable in Indonesian, Khmer, Tagalog, Thai and Vietnamese. Perhaps this entire aspect of the project was quixotic from the very beginning. Recall that the process began with the creation of English language prototypes for the listening test, oral interview, dictation and cloze procedure. These prototypes were then translated into the five languages with strict instructions to really translate them, ie. to make them comfortably and wholly Indonesian, Khmer, Tagalog, Thai and Vietnamese. While the very act of translating the passages in five different directions probably affected their comparability across languages, they probably remained at least roughly the same at this stage of development. Then, during the summer of 1988, the tests were administered, analyzed and revised separately using different samples of students with the result that the tests further diverged in content and function.

We now know that the use of English language prototypes for the development of these tests may have created problems that we did not foresee. One danger is that such a strategv avoids the use of language that is authentic in the target language. For instance, a passage that is translated from English for use in Khmer cloze test may be topic that would never be discussed in the target culture, may be organized in a manner totally alien to Khmer, or may simply seem stilted to native speakers of Khmer because of its rhetorical structure. These problems could occur no matter how well-translated the passage might be.

Ultimately, the tests did not turn out to be similar enough across languages to justify using this translation strategy. Thus we do not recommend its use in further test development projects. It would probably have been far more profitable to use authentic materials from the countries involved to develop tests directly related to the target languages and cultures.

CONCLUSION

In summary, the tests in each of the five SEASSI Proficiency Examinations appear to be reasonably well-centered and seem to adequately disperse the students' performance. They are also reasonably reliable. Naturally, future research should focus on ways to make the tests increasingly reliable and further build a case for their validity. Thus the final versions of the tests can be passed on to future SEASSIs at other sites with some confidence that any decisions based on them will be reasonably professional and sound. It is also with some confidence that the tests will be used here at the University of Hawaii at Manoa to test the overall proficiency of students studying Indonesian, Khmer, Tagalog, Thai and Vietnamese. However, the process of test development and revision should never be viewed as finished. Any test can be further improved and made to better serve the population of students and teachers who are the ultimate users of such materials.

One final point must be stressed: we could never have successfully carried out this project without the cooperation of the many language teachers who volunteered their time while carrying out other duties in the Indo-Pacific Languages department, or the SEASSIs held at University of Hawaii at Manoa. We owe each of these language teachers a personal debt of gratitude. Unfortunately, we can only thank them as a group for their professionalism and hard work.

REFERENCES

*ACTFL (1986). **ACTFL proficiency guidelines.** Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.*

*Anderson-Bell. (1989). **ABSTAT.** Parker, CO: Anderson-Bell.*

*APA. (1985). **Standards for educational and psychological testing.** Washington, DC: American Psychological Association.*

*Bachman, L & S Savignon. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. **Modern Language Journal,** 70, 380-397.*

*Borland. (1989). **QuattroPro.** Scotts Valley, CA: Borland International.*

*Brown, J D. (1983). A closer look at cloze: validity and reliability. In J W Oller, Jr (Ed). **Issues in language testing.** Cambridge, MA: Newbury House.*

*Brown, J D. (1984). A cloze is a cloze is a cloze? In J Handscombe, R A Orem, and B P Taylor (Eds). **On TESOL '83: the question of control.** Washington, DC: TESOL.*

*Brown, J D. (1988a). **Understanding research in second language learning: A teacher's guide to statistics and research design.** London: Cambridge University.*

*Brown, J D. (1988b). Tailored cloze: improved with classical item analysis techniques. **Language Testing,** 5, 19-31.*

Brown, J D., H G Cook, C Lockhart and T Ramos (Unpublished ms). The SEASSI Proficiency Examination Technical Manual. Honolulu, HI: University of Hawaii at Manoa.

Carroll, J B and S M Sapon (1959). ***Modern language aptitude test.*** *New York: The Psychological Corporation.*

Cronbach, L J. (1970). ***Essentials of psychological testing*** *(3rd ed). New York: Harper and Row.*

Ebel, R L. (1979). ***Essentials of educational measurement*** *(3rd ed). Englewood Cliffs, NJ: Prentice-Hall.*

Fry, E B. (1976). ***Fry readability scale (extended).*** *Providence, RI: Jamestown Publishers.*

ILR (1982). Interagency Language Roundtable Language Skill Level Descriptions: Speaking. appendix B in Liskin-Gasparro (1982).

Kuder, G F and M W Richardson. (1937). The theory of estimation of test reliability. ***Psychometrika,*** *2, 151-160.*

Liskin-Gasparro (1982). ***Testing and teaching for oral proficiency.*** *Boston: Heinle and Heinle.*

Savignon, S J. (1985). Evaluation of communicative competence: the ACTFL provisional proficiency guidelines. ***Modern Language Journal,*** *69, 129-142.*

Wilkinson, L. (1988). ***SYSTAT: The system for statistics.*** *Evanston, IL: SYSTAT.*

APPENDIX A
ACTFL PROFICIENCY GUIDELINES FOR SPEAKING AND LISTENING
(ACTFL 1986)

Generic Descriptions-Speaking

Novice The Novice level is characterized by the ability to communicate minimally with learned material.

Novice Low Oral production consists of isolated words and perhaps a few high frequency phrases. Essentially no functional communicative ability.

Novice Mid Oral production continues to consist of isolated words and learned phrases within very predictable areas of need, although quantity is increased. Vocabulary is sufficient only for handling simple, elementary needs and expressing basic courtesies. Utterances rarely consist of more than two or three words and show frequent long pauses and repetition of interlocutor's words. Speaker may have some difficulty producing even the simplest utterances. Some Novice-Mid speakers will be understood only with great difficulty.

Novice-High Able to satisfy partially the requirements of basic communicative exchanges by relying heavily on learned utterances but occasionally expanding these through simple recombinations of their elements. Can ask questions or make statements involving learned material. Shows signs of spontaneity although this falls short of real autonomy of expression. Speech continues to consist of learned utterances rather than of personalized, situationally adapted ones. Vocabulary centers on areas such as basic objects, places, and most common kinship terms. Pronunciation may still be strongly influenced by first language. Errors are frequent and, in spite of repetition, some Novice-High speakers will have difficulty being understood even by sympathetic interlocutors.

Intermediate The Intermediate level is characterized by the speaker's ability to:
—create with the language by combining and recombining learned elements, though primarily in a reactive mode,
—initiate, minimally sustain, and close in a simple way basic communicative tasks, and
—ask and answer questions.

Intermediate-Low Able to handle successfully a limited number of interactive, task-oriented and social situations. Can ask and answer questions, initiate and respond to simple statements and maintain face-to-face conversation, although in a highly restricted manner and with much linguistic inaccuracy. Within these limitations, can perform such tasks as introducing self, ordering a meal, asking directions, and making purchases. Vocabulary is adequate to express only the most elementary needs. Strong interference from native language may occur. Misunderstandings frequently arise, but with repetition, the Intermediate-Low speaker can generally be understood by sympathetic interlocutors.

Intermediate Mid Able to handle successfully a variety of uncomplicated, basic and communicative tasks and social situations. Can talk simply about self and family members. Can ask and answer questions and participate in simple conversations on topics beyond the most immediate needs; e.g., personal history and leisure time activities. Utterance length increases slightly, but speech may continue to be characterized by frequent long pauses, since the smooth incorporation of even basic conversational strategies is often hindered as the speaker struggles to create appropriate language forms. Pronunciation may continue to be strongly influenced by first language and fluency may still be strained. Although misunderstandings still arise, the Intermediate-Mid speaker can generally be understood by sympathetic interlocutors.

Intermediate High Able to handle successfully most uncomplicated communicative tasks and social situations. Can initiate, sustain, and close a general conversation with a number of strategies appropriate to a range of circumstances and topics, but errors are evident. Limited vocabulary still necessitates hesitation and may bring about slightly unexpected circumlocution. There is emerging evidence of connected discourse, particularly for simple narration and/or description. The Intermediate-High speaker can generally by understood even by interlocutors not accustomed to dealing with speakers at this level, but repetition may still be required.

Advanced

The Advanced level is characterized by the speaker's ability to:
- converse in a clearly participatory fashion;
- initiate, sustain, and bring to closure a wide variety of communicative tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events;
- satisfy the requirements of school and work situations; and
- narrate and describe with paragraph-length connected discourse

Advanced

Able to satisfy the requirements of everyday situations and routine school and work requirements. Can handle with confidence but not with facility complicated tasks and social situations, such as elaborating, complaining, and apologizing. Can narrate and describe with some details, linking sentences together smoothly. Can communicate facts and talk casually about topics of current public and personal interest, using general vocabulary. Shortcomings can often be smoothed over by communicative strategies, such as pause fillers, stalling devices, and different rates of speech. Circumlocution which arises from vocabulary or syntactic limitations very often is quite successful, though some groping for words may still be evident. The Advanced level speaker can be understood without difficulty by native interlocutors.

Advanced Plus

Able to satisfy the requirements of a broad variety of everyday, school, and work situations. Can discuss concrete topics relating to particular interests and special fields of competence. There is emerging evidence of ability to support opinions, explain in detail, and hypothesize. The Advanced Plus speaker often shows a well developed ability to compensate for an imperfect grasp of some forms with confident use of communicative strategies, such as paraphrasing and circumlocution. Differentiated vocabulary and intonation are effectively used to communicate fine shades of meaning. The Advanced-Plus speaker often shows remarkable fluency and ease of speech but under the demands of Superior level, complex tasks, language may break down or prove inadequate

Superior

The Superior level is characterized by the speaker's ability to
- participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics; and
- support opinions and hypothesize using native like discourse strategies

Superior

Able to speak the language with sufficient accuracy to participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics. Can discuss special fields of competence and interest with ease. Can support opinions and hypothesize, but may not be able to tailor language to audience or discuss in depth highly abstract or unfamiliar topics. Usually the Superior level speaker is only partially familiar with regional or other dialectical variants. The Superior level speaker commands a wide variety of interactive strategies and shows good awareness of discourse strategies. The latter involves the ability to distinguish main ideas from supporting information through syntactic, lexical and suprasegmental features (pitch, stress, intonation). Sporadic errors may occur, particularly in low-frequency structures and some complex high-frequency structures more common to formal writing, but no patterns of error are evident. Errors do not disturb the native speaker or interfere with communication

Generic Descriptions-Listening

These guidelines assume that all listening tasks take place in an authentic environment at a normal rate of speech using standard or near standard norms

Novice Low

Understanding is limited to occasional isolated words, such as cognates, borrowed words, and high frequency social conventions. Essentially no ability to comprehend even short utterances.

Novice Mid

Able to understand some short, learned utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends some words and phrases from simple questions, statements, high-frequency commands and courtesy formulae about topics that refer to basic personal information or the immediate physical setting. The listener requires long pauses for assimilation and periodically requests repetition and/or a slower rate of speech.

Novice High — Able to understand short, learned utterances and some sentence length utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends words and phrases from simple questions, statements, high frequency commands and courtesy formulae. May require repetition, rephrasing and/or a slowed rate of speech for comprehension

Intermediate Low — Able to understand sentence length utterances which consist of recombinations of learned elements in a limited number of content areas, particularly if strongly supported by the situational context. Content refers to basic personal background and needs, social conventions and routine tasks, such as getting meals and receiving simple instructions and directions. Listening tasks pertain primarily to spontaneous face to face conversations. Understanding is often uneven; repetition and rewording may be necessary. Misunderstandings in both main ideas and details arise frequently

Intermediate Mid — Able to understand sentence length utterances which consist of recombinations of learned utterances on a variety of topics. Content continues to refer primarily to basic personal background and needs, social conventions and somewhat more complex tasks, such as lodging, transportation, and shopping. Additional content areas include some personal interests and activities, and a greater diversity of instructions and directions. Listening tasks not only pertain to spontaneous face-to-face conversations but also to short routine telephone conversations and some deliberate speech, such as simple announcements and reports over the media. Understanding continues to be uneven.

Intermediate High — Able to sustain understanding over longer stretches of connected discourse on a number of topics pertaining to different times and places; however, understanding is inconsistent due to failure to grasp main ideas and/or details. Thus, while topics do not differ significantly from those of an Advanced level listener, comprehension is less in quantity and poorer in quality

Advanced — Able to understand main ideas and most details of connected discourse on a variety of topics beyond the immediacy of the situation. Comprehension may be uneven due to a variety of linguistic and extralinguistic factors, among which topic familiarity is very prominent. These texts frequently involve description and narration in different time frames or aspects, such as present, nonpast, habitual, or imperfective. Texts may include interviews, short lectures on familiar topics, and news items and reports primarily dealing with factual information. Listener is aware of cohesive devices but may not be able to use them to follow the sequence of thought in an oral text.

Advanced Plus — Able to understand the main ideas of most speech in a standard dialect; however, the listener may not be able to sustain comprehension in extended discourse which is propositionally and linguistically complex. Listener shows an emerging awareness of culturally implied meanings beyond the surface meanings of the text but may fail to grasp sociocultural nuances of the message

Superior — Able to understand the main ideas of all speech in a standard dialect, including technical discussion in a field of specialization. Can follow the essentials of extended discourse which is propositionally and linguistically complex, as in academic/professional settings, in lectures, speeches, and reports. Listener shows some appreciation of aesthetic norms of target language, of idioms, colloquialisms, and register shifting. Able to make inferences within the cultural framework of the target language. Understanding is aided by an awareness of the underlying organizational structure of the oral text and includes sensitivity for its social and cultural references and its affective overtones. Rarely misunderstands but may not understand excessively rapid, highly colloquial speech or speech that has strong cultural references

Distinguished — Able to understand all forms and styles of speech pertinent to personal, social and professional needs tailored to different audiences. Shows strong sensitivity to social and cultural references and aesthetic norms by processing language from within the cultural framework. Texts include theater plays, screen productions, editorials, symposia, academic debates, public policy statements, literary readings, and most jokes and puns. May have difficulty with some dialects and slang