On Developing an Oral Proficiency Test for English as a Foreign Language

著者（英）	Elizabeth Hiser
journal or publication title	Kansai Gaidai Educational Research and Report
volume	2
page range	15-27
year	2001-08
URL	http://id.nii.ac.jp/1443/00005685/

関西外国語大学 教育研究報告 第2号（2001年8月）
Kansai Gaidai Educational Research and Report, No.2 (Aug. 2001)

On Developing an Oral Proficiency Test for English as a Foreign Language

Elizabeth Hiser

Abstract

This paper illustrates the process of developing an oral proficiency test using a procedure taken from Carroll and Hall's (1985) practical guide to making language tests. The development generally follows the steps given, is piloted and then used in placing students in a four year English language program for international communication in Japan. The rationale for the structure of the test and guidelines for administering it are described in detail. The actual test that was developed proved to be not only valid and highly reliable but also more accurate than either the TOEFL or the Michigan Placement Test, Form A, in determining the communicative English language ability of the sample. Statistical analyses were used to support this argument and conclusion.

Introduction

As the Ministry of Education directives on communicative methodology hit the classroom in Japan, the need for and the value of an oral proficiency test that is simple, efficient, effective, reliable and reasonably priced in terms of time and material expenditures increases. In response to just such a situation, an oral proficiency test was developed for use in placement within a graded English language program at the tertiary level of education.

Historically, there are excellent choices available (Alderson, Krahnke, & Stansfield, 1987; Kitao & Kitao, 1999; Banerjee, Clapham, Clapham, & Wall, 1999); but descriptively, they are usually lacking in one of the adjectives listed above – simple, efficient, effective, or reliable. The Oral Proficiency Test (OPT) described below was developed specifically with these criter-

ia in mind. Carroll and Hall's (1985) steps in designing an oral test were generally followed with modification in procedure as required. Those five steps are listed in Table 1 with related descriptive notes.

Table 1: Carroll and Hall's Steps for Designing an Oral Test.

1. Design & Planning
   a) checking the sample needs; who is to be tested and why.
   b) investigating the test setting and equipment available.
   c) listing the items and points to be tested.
   d) selecting the format best suited for testing these points.
2. Development, Piloting and Preparation
   a) writing and developing the test items, passages, procedures or answer sheets.
   b) developing student instructions.
   c) piloting the test with a comparable sample.
   d) rewriting, correcting the pilot problems.
   e) reproducing the test for use.
3. Operationalizing & Administering the Test
   a) note any additional problematic areas on the test or in procedures.
   b) score the test.
   c) record data from the results.
4. Analysize the Results
   a) check reliability.
   b) do an item analysis.
   c) do a factorization of items.
   d) check the distribution of scores.
   e) compare the results with other criteria.
   f) do other analyses that may be relevant, such as t-tests or multiple regression.
5. Monitor the Test
   a) establish test security procedures for administration and storage.
   b) continue monitoring the effectiveness of the test.
   c) create alternative forms of the test and verify the reliability.
   d) establish the equivalency of the forms.

The first issue in designing a language proficiency test of any kind that must be addressed is whether one views language acquisition as holistic or componential. There is no doubt the line of acquisition is more or less graphed from lower left to upper right, within a band of weaknesses and strengths that any particular individual may possess along the way. If acquisition is componential in nature what areas or components need to be evaluated in order to determine oral proficiency? In this study we view language acquisition to be a process of acquiring overlapping components of skills; related to a limited extent but generally independent in nature. Those components determined to contribute to second language speaking ability or oral proficiency in-

clude fluency, pronunciation, vocabulary use, syntax, delivery, and content of message. There can be no doubt cognitive ability contributes to oral proficiency at least at the level of organization, along with delivery elements such as speed of processing, gestures and/or eye contact.

Procedure

In responding to the first item of Carroll and Hall's steps in the creation of an oral test, it was determined that the students to be tested would range in ability thus requiring an open scale for rating, since placement in a streamed program was the objective. Following the second item of Carroll and Hall's work suggested that since native speakers were available to administer the test, a possible interview format was the best suited means of determining the individual levels of students as long as the interviewers were given a standardized format with specific criteria to be evaluated for the interview. The components as described above were six independent scales representing 1) content, organization and interest, 2) delivery including eye contact and self-expression, 3) grammaticality and structure use, 4) vocabulary, 5) pronunciation, and 6) fluency or pace of presentation. With the exception of grammaticality, checking these points does not rely on the use of discreet items, so that the structure of the interview could be free of the detailed procedures developed on other more complex or sophisticated oral testing procedures such as the Oral Proficiency Interview (ETS, 1984) or the Speaking Proficiency English Assessment Kit (ETS, 1992) used elsewhere.

Table 2: Component Scales for the Oral Proficiency Test

Content:

5 Excellent; well thought out and interesting.
4 Good; interesting but not in the best of order.
3 OK; interest level acceptable but disorganized.
2 Poor; little relevance, order or interest.
1 Unsatisfactory; content vague or questionable.
0 No score.

Delivery:

5 Confident and clear, good eye contact.
4 Clear but slightly nervous, some eye contact.
3 Hesitant but capable of expressing self, ideas.
2 Nervous and hesitant, difficult to follow.
1 Hard or impossible to follow.
0 No score.

Grammar:

5 A few or no minor mistakes.
4 Several minor mistakes.
3 Structural mistakes from time to time.
2 Noticeable mistakes throughout.
1 Uses Japanese sentence structure for the most part.
0 No score.

Vocabulary:

5 High, words used correctly.
4 Good, some word choice not the best.
3 Common words used, repetitive.
2 Word choice not often correct.
1 Lacking in word choice and correctness.
0 No score.

Pronunciation:

5 Excellent; no problems.
4 Good, but a few minor mistakes.
3 Obvious Japanese b/v, l/r, s/sh, etc., mistakes.
2 Intonation and stress inappropriate, but understandable.
1 Incomprehensible.
0 No score.

Fluency:

5 Excellent speed, eye contact.
4 Good; though some hesitancy, and slowness or responses.
3 Stutters, halts to find the right words or structure.
2 Minimal eye contact and much "thinking" or long pauses.
1 No confidence, extreme nervousness, minimal responses.
0 No score.

At Step Two of the test development process, the scales for each component were developed into a formal score sheet for interviewers to use. See Table 2 for the description of each component as delineated. How best to elicit language from students for assessment was then carefully considered.

It was decided during the process of determining procedures, that emphasis should fall upon the student being "interviewed" to produce language rather than the student's response to the interviewer's elicitation of language. This point is often the most difficult in this assessment format as the interviewer needs to maintain a mental balance between listening to the student and evaluating them. If the student is required to produce language rather than be constantly prompted in a structured but artificial conversation, then the interviewer is free to focus on assessment of the language production without worrying about interacting with the testee. With this issue in mind, it was further decided that students would be required to speak upon a

selected topic for a brief period of time during which language proficiency assessment could be attempted. Since this was to be an impromptu speech as it were, several topics would be provided for the testee to choose from allowing them the opportunity to select a topic upon which they felt they could perform best, and avoiding the possible problem of examinees drawing a blank and not being able to think of a topic themselves. This approach would permit the person doing the evaluation to simply concentrate on the assessment and circle the rate or score (on a prepared evaluation sheet) which they felt best applied to the language produced.

When selecting topics for use with the test, some attempt was made to provide a broad band of subjects, but also to make them personal enough that details would not require expertise in areas that might be unfamiliar to the examinees. These topics were offered as suggestions and students were not limited to use of this list. Consideration for what areas or topics students might have the best cognitive schema was given. See Table 3 for a brief list of the topics offered. These were presented in random groups of approximately five or six to the examinees with alternatives readily available if students could not comfortably select one from the group initially provided.

Table 3: Suggested Topics Presented to Examinees for the OPT

Buses	Your Room	Your Favorite Food	Festivals
Exercise or Sports	Portable Telephones	Health and Beauty	Cars and Driving
Studying English	Your Favorite City	Hot Spring Resorts	Your Favorite Song or Music
Your Hobby	High School	Trains	Your Hometown
Your Favorite Film or Movie	Your Travel and Trips	A Famous Person	Your Favorite Actor or Actress
Smoking	Your Family	Bicycles	Vacations
Your Future Plans	Hiking and Camping	Your Favorite Book	A Foreign Country Visited
Fishing	Guns and Knives	Your Pet	Shopping
Shoes	Boats	Christmas Time	Fashion
A Teacher You Liked	Dangerous Things To Do	Your Favorite Season	Comic Books
Jewelry	Animation	Watches and Clocks	Cooking
Your Favorite Book	New Year's Time	Your Favorite Restaurant	Your Part-time Job
Money	Photographs	Golden Week	Karaoke
Fast Food	Free Time	Lunch	The News

The instrument created by these decisions was then piloted at a commercial language school in Kyoto where proficiency is regularly monitored by use of the Michigan English Placement Test (MEPT). It was found that the interview evaluations from the Oral Proficiency Test (OPT) matched the ranking of students by the MEPT Form C perfectly for a correlation between scores of 1.00 ($\underline{N}$ = 20).

Encouraged by this success, arrangements were made to administer the OPT to the incoming freshman class of a four year university program in international language and communication. Ten native English speaking language instructors were selected to give the test to 164 students. The final streaming or tracking by English language ability for the program was to be based on a combination of the students' TOEFL scores and the OPT assessments. Each instructor was randomly assigned between 15 and 18 students for evaluation with the understanding that the assessment would probably only require 10~15 minutes per student.

Test Administration Guidelines

For the sake of consistency in administering the OPT, guidelines and instructions were prepared for the faculty giving the test, and a general meeting of those involved was called at which testing procedures were discussed in advance. The guidelines included five major points. First, the type of assessment being undertaken was to place responsibility on the examinee to speak and not the examiner to prompt and direct conversation. It was not to be conversational in nature and the students were to be the focus of the session. In that respect this was not to be an interview but a simple speaking evaluation or oral proficiency test.

Secondly, after greeting the students and offering them a seat, perhaps helping them to relax, the student was to be given one of the compiled lists of topics from which to choose a subject for their presentation. During this time the test administrator was to write the student's name, ID number and date at the top of the evaluation form provided.

Third, the student was to be given one to two minutes to consider and prepare a short statement on the chosen topic. If the student had any difficulty in choosing a topic from the list, they were to be given a new list to consider. Examinees were then expected to speak for three to five minutes on the topic they had chosen while the examiner listened and scored the appropriate marks on the evaluation form. Closing comments should include thanking the students and asking them to send in the next candidate. Totals for the six sections on the test were then to be made and entered at the top of the evaluation form.

Finally, instructors were told that if a student was unable to speak at length on the topic

chosen, they might feel free to ask a question to encourage further development of the topic, but if the student again failed to speak adequately on the subject, examiners should attempt to evaluate them as best possible and let them go.

Warnings to the examiners were then made concerning the evaluation procedures. They included not accepting memorized speeches that students might have prepared in advance, and not accepting the use of notes or reading. Neither were examiners to allow students to draw them into talking or interacting with them to any great extent. The instructors were encouraged not to show the evaluation form to the students and to total the scores only when the examinee had left the room. It was also suggested that the examiners change the topic list often so that students waiting outside understood that the topics were changing and would not attempt to prepare in advance for the assessment, i.e. examiners should help avoid possible washback. Finally, it was made clear that the scales on the evaluation form were considered relatively independent and that students were not to be evaluated the same on all scales unless they were actually the same level in each category. In other words, a student might have excellent pronunciation, but not interesting content. Likewise, they might have their thoughts well-organized, but not have adequate vocabulary to express those ideas. Each component of the test was to be evaluated as independently as possible.

Results

Step Four of Carroll and Hall's procedure calls for analyzing the data from the results obtained in administering the test. To this end, the basic Gaussian curve of scores was initially examined. See Figure 1 for the distribution of those scores. The overall shape of the curve appears to be bi-peaked with some skewedness to the right. The higher end of the curve appears heavier than expected, and may indicate some test administrators were too easy in their evaluations of the examinees. The mean score for the sample on the test was 20, with a standard deviation of 5.98 ($\underline{N}$ = 164). A brief examination of the mean scores for each instructor's group of students does show significant differences (Probability of F = .0310) between their mean scores, but it is also possible that they were assigned clusters of students of differing ability. See Table 4 for the analysis of variance and descriptions of subpopulations assigned each instructor. These results cannot be taken as proof that some examiners scored better than others, but it does point in that direction.

Reliability of the scores was considered next to determine the accuracy of the test and the consistency of the component scores within the total. Cronbach's alpha, a conservative estimate

Figure 1: Histogram of OPT Scores, N = 164

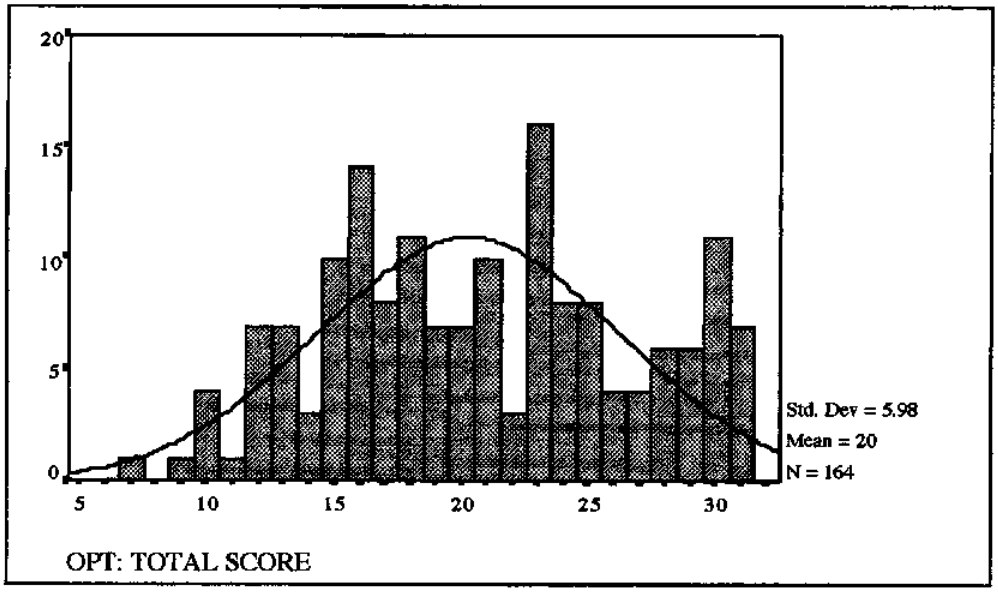

Table 4: Analysis of Variance and Description of Subpopulations for Each Test Administrator's Group of Students

Analysis of Variance

Variable TOTAL OPT SCORE By Variable EXAMINER

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	9	642.0450	71.3383	2.1187	.0310
Within Groups	154	5185.1989	33.6701		
Total	163	5827.2439			

Description of Subpopulations

Summaries of TOTAL OPT SCORE By levels of EXAMINER

Variable	Value	Mean	Std Dev	Cases
	For Entire Population:	19.7439	5.9791	164
EXAMINER	1	18.2500	6.2557	16
EXAMINER	2	23.3750	6.1631	16
EXAMINER	3	20.6471	6.9637	17
EXAMINER	4	18.4000	3.9964	15
EXAMINER	5	22.1875	6.0357	16
EXAMINER	6	19.4444	5.7622	18
EXAMINER	7	18.7059	6.6686	17
EXAMINER	8	17.5882	5.5684	17
EXAMINER	9	21.6875	5.0295	16
EXAMINER	10	17.2500	4.6975	16

(Youngman, 1979), was determined to be .9472 for the instrument, providing an unexpected validation of the componential structure of the OPT. See Table 5 for the Scale Mean, the Scale Variance, the Corrected Item-Total Correlation, and the Alpha if Item Deleted figures for this analysis. This high value indicates the scores were generally the same for all components of the test which shows a consistency of scores for individuals in each of the areas evaluated. The argument then being that these elements of language acquisition improve inter-dependently as overall language acquisition improves, rather than as independent components, unrelated to overall improvement.

Table 5: Reliablility Analysis

RELIABLITY ANALYSIS - SCALE

Statistics for	Mean	Variance	Std Dev	Variables
SCALE	19.8811	35.8769	5.9897	6

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Correctd Item-Total Correlation	Alpha if Item Deleted
CONTENT	16.5213	25.2802	.8525	.9356
DELIVERY	16.6341	24.3224	.8823	.9319
FLUENCY	16.6890	24.4242	.8385	.9376
GRAMMAR	16.6189	24.4505	.8438	.9368
PRONUN	16.3628	27.6237	.7493	.9476
VOCAB	16.5793	24.9661	.8767	.9327

Reliability Coefficients
N of Cases = 164.0 N of Items = 6
Alpha = .9472

Additional support for this argument can be gained from the correlations among the components. See Table 6 for the Spearman's rho value among the variables for the bivariate correlations. The lowest value in the table is .6081, a moderate value. The other values are even stronger, considering the test administrators were specifically instructed to consider each component independently. There is also very high statistical significance for all correlations (P = .000).

Multiple regression analysis was computed to determine which of the components best predicted the total score for the test. *Delivery* was accepted first in a stepwise analysis producing a Multiple R of .91803. In fact, all components were selected to enter the equation at some point showing each contributed to the calculation of the equation in predicting the total score on the OPT. See Table 7 for the order in which the variables were accepted into the equation, the R^2 value for each, and the final constant.

Table 6: Spearman Two-tailed Correlations Among OPT Components

DELIVERY	.8419				
	N(164)				
	Sig .000				
FLUENCY	.8081	.8465			
	N(164)	N(164)			
	Sig .000	Sig .000			
GRAMMAR	.7418	.7737	.7248		
	N(164)	N(164)	N(164)		
	Sig .000	Sig .000	Sig .000		
PRONUN	.6354	.6383	.6081	.7570	
	N(164)	N(164)	N(164)	N(164)	
	Sig .000	Sig .000	Sig .000	Sig .000	
VOCAB	.7813	.8181	.7979	.8185	.7413
	N(164)	N(164)	N(164)	N(164)	N(164)
	Sig .000	Sig .000	Sig .000	Sig .000	Sig .000
	CONTENT	DELIVERY	FLUENCY	GRAMMAR	PRONUN

Table 7: Multiple Regression (Step Six) Predicting the Total OPT Score

Step	Variable	B	SEB	Beta	T	Sig T	Multiple R	R^2
1	DELIVERY	.811322	.106887	.158302	7.590	.0000	.91039	.82881
2	VOCABULARY	.993810	.104070	.183796	9.549	.0000	.95907	.91982
3	GRAMMAR	.985307	.090901	.197330	10.839	.0000	.97338	.94747
4	CONTENT	1.056565	.100561	.193661	10.507	.0000	.98307	.96642
5	FLUENCY	1.098519	.091720	.221632	11.977	.0000	.98886	.97731
6	PRONUN	1.016535	.095932	.159212	10.596	.0000	.99336	.98677
	(Constant)	-.018896	.221348		-.085	.9321		

Additional attention was given to the issue of validity. In an attempt to demonstrate criterion validity, correlations with the TOEFL scores and Michigan Placement Test (Form A) scores for each student were calculated. See Table 8 for the results. These correlations included the scores for each section of the tests – listening (MLISTOT), reading (MREADTOT), vocabulary (MVOCTOT), and structures (MGRMTOT) for the Michigan, Form A; listening (TOEFLIS), reading (TOEFLRDG), and structures (TOEFLSTR) for the TOEFL; and delivery (DELIVERY), vocabulary (DEVOCAB), fluency (DFLUENCY), structures (DGRAMMAR), pronunciation (DPRONUN), and content (DCONTENT) for the OPT. Totals for both the MEPT-A and the OPT were also entered into the matrix; labeled as variables ZMEPTOT for the MEPT-A and ZOPTOTAL for the OPT.

Table 8: Correlations Among OPT, TOEFL and MPT-A for Students

TOEFLS	.0035	.2636	.0257	.0480				
	N(162)	N(162)	N(162)	N(162)				
	Sig .965	Sig .001	Sig .745	Sig .544				
TOEFLRDG	.0325	-.1024	.0488	.1355	.3249			
	N(162)	N(162)	N(162)	N(162)	N(162)			
	Sig .682	Sig .195	Sig .538	Sig .085	Sig .000			
TOEFLSTR	.0617	-.1066	.1664	.1249	.2586	.5388		
	N(162)	N(162)	N(162)	N(162)	N(162)	N(162)		
	Sig .435	Sig .177	Sig .034	Sig .113	Sig .001	Sig .000		
ZMEPTOT	.6404	.4410	.6279	.5092	.2630	.1188	.2478	
	N(69)	N(69)	N(69)	N(69)	N(64)	N(64)	N(64)	
	Sig .000	Sig .000	Sig .000	Sig .000	Sig .036	Sig .350	Sig .048	
ZOPTOTAL	.0033	.1036	-.0326	.1088	.5898	.2213	.2516	.0878
	N(164)	N(164)	N(164)	N(164)	N(161)	N(161)	N(161)	N(63)
	Sig .966	Sig .187	Sig .678	Sig .165	Sig .000	Sig .005	Sig .001	Sig .494
DCONTENT	-.0028	.0416	.0135	.0468	.5474	.2073	.2510	.0363
	N(164)	N(164)	N(164)	N(164)	N(161)	N(161)	N(161)	N(63)
	Sig .972	Sig .597	Sig .864	Sig .552	Sig .000	Sig .008	Sig .001	Sig .778
DELIVERY	-.0596	.1032	.0178	.0805	.5646	.1784	.1951	-.0156
	N(164)	N(164)	N(164)	N(164)	N(161)	N(161)	N(161)	N(63)
	Sig .448	Sig .188	Sig .821	Sig .305	Sig .000	Sig .024	Sig .013	Sig .903
DEVOCAB	.0314	.0768	-.0552	.1317	.5106	.2356	.2514	.0797
	N(164)	N(164)	N(164)	N(164)	N(161)	N(161)	N(161)	N(63)
	Sig .689	Sig .328	Sig .482	Sig .093	Sig .000	Sig .003	Sig .001	Sig .535
DFLUENCY	-.0307	.0692	-.0493	.0572	.5066	.1878	.2175	.0282
	N(164)	N(164)	N(164)	N(164)	N(161)	N(161)	N(161)	N(63)
	Sig .697	Sig .378	Sig .530	Sig .467	Sig .000	Sig .017	Sig .006	Sig .826
DGRAMMAR	.0042	.0906	-.0378	.1097	.5503	.2656	.2546	.1659
	N(164)	N(164)	N(164)	N(164)	N(161)	N(161)	N(161)	N(63)
	Sig .958	Sig .249	Sig .630	Sig .162	Sig .000	Sig .001	Sig .001	Sig .194
DPRONUN	.0718	.1141	-.0004	.1500	.4479	.1504	.2140	.2297
	N(164)	N(164)	N(164)	N(164)	N(161)	N(161)	N(161)	N(63)
	Sig .361	Sig .146	Sig .996	Sig .055	Sig .000	Sig .057	Sig .006	Sig .070
	MGRMTOT	MLISTOT	MREADTOT	MVOCTOT	TOEFLIS	TOEFLRDG	TOEFLSTR	ZMEPTOT

Correlation of these 15 variables shows low to moderate coefficients between both the total scores on the MEPT-A and the OPT for the TOEFL components listening, reading and structures (MEPT-A: .2630, .1188, .2478, and OPT: .5898, .2213, .2516). All of these are highly significant except the lowest (.1188) between the MEPT-A and TOEFL reading. The OPT per-

formed better than the MEPT-A in every case. Additionally the six components of the OPT all correlated moderately with the TOEFL listening score with very high significance (content: .5475, delivery: .5646, vocabulary: .5106, fluency: .5066, grammar: .5503, and pronunciation: .4479). There were also low but highly significant coefficients for each of the OPT components and the other two sections of the TOEFL – reading and structures.

The final analysis enlisted in the investigation of validity was factorization. The 13 component scores from the TOEFL, OPT, and MEPT-A were factored into two groups based on the scree plot and initial results from preliminary work. See Table 9 for the item loading on each of the factors and the division of the components into two groups. The three TOEFL components loaded most strongly on the first factor made up of the OPT scores, but also loaded as complex items on the second factor which is made up of MEPT-A scores.

Table 9: Factorization of OPT Component Scores, TOEFL Scores, and MEPT-A Component Scores

Pairwise deletion of cases with missing values. Extraction 1 for analysis 1, Principal Components Analysis (PC). PC extracted 2 factors. VARIMAX rotation 1 for extraction. 1 in analysis 1 – Kaiser Normalization. VARIMAX converged in 3 iterations.

Rotated Factor Matrix:

	Factor 1	Factor 2
DELIVERY	.91371	
DEVOCAB	.90513	
DCONTENT	.89214	
DGRAMMAR	.89054	
DFLUENCY	.88323	
DPRONUN	.79919	
TOEFLIS	.60488	.25508
TOEFLSTR	.30927	.28309
TOEFLRDG	.28355	.21241
MGRMTOT		.75900
MVOCTOT		.73380
MREADTOT		.62933
MLISTOT		.54131

The first factor is, of course, the primary source of variance in the analysis, and in this case the initial solution showed the first six items, components of the OPT to account for 84.9% of the entire variance, more than the TOEFL components contributed. This would indicate the OPT provided most of the variance in scores and was therefore the most accurate measure in discriminating among the students in ability.

Conclusion

This paper has illustrated the process of developing an oral proficiency test using Carroll and Hall's procedure. The actual test that was developed proved to be not only valid and highly reliable but also more accurate than either the TOEFL or the Michigan Placement Test, Form A, in determining the communicative ability of the sample. Statistical analyses were used to support this argument and conclusion.

As for the fifth and final step in the Carroll and Hall procedure, the OPT is available for use with other samples and further research should include recording the interview sessions so that inter-rater reliability might also be confirmed.

References

Alderson, J. C., Krahnke, K. J. & Stansfield, C. W. (Eds.) (1982). *Reviews of English Language Proficiency Tests*. Washington, D. C.: Teachers of English to Speakers of Other Languages.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

Banerjee, J., Clapham, C., Clapham, P. & Wall, D. (Eds.) (1999). *ILTA Language Testing Bibliography 1990-1999*. Lancaster: Department of Linguistics and Modern English Language, Lancaster University.

Carroll, B. J. & Hall, P. J. (1985). *Make Your Own Language Tests: A Practical Guide to Writing Language Performance Tests*. London: Pergamon Institute of English.

Carroll, B. J. (1984). *Testing Communicative Performance*. London: Pergamon Institute of English.

Delarche, M. & Marshall, N. (1996). *Communicative Oral Testing* in van Troyer, G., Cornwell, S. & Morikawa, H. (Eds.) *On JALT 95: Curriculum & Evaluation*. (Proceedings of the JALT 1995 International Conference) Tokyo: Japan Association of Language Teachers.

ETS (1984). *Oral Proficiency Test*. Princeton, NJ: Educational Testing Service.

ETS (1992). *Speaking Proficiency English Assessment Kit (SPEAK)*. Princeton, NJ: Educational Testing Service.

Kitao, K. & K. Kitao, (1999). *Language Testing*. [On-line]. Available: <http://ilc2.doshisha.ac.jp/users/kkitao/online/www/test.html>

Madsen, H. S. (1983). *Techniques in Testing*. Oxford, UK: Oxford University Press.

Youngman, M. (1979). *Analysing Social and Educational Research Data*. London: McGraw-Hill Book Company (UK) Limited.