Testing Communicative Competence and Performance in UMSEP

Khong Chooi Peng and Subramaniam Rajagopal
Pusat Bahasa
Universiti Malaya

Introduction

Historically, language testing has tended to lag behind developments in language teaching pedagogy This state of affairs has continued to the present resulting in what Morrow (1979:143) describes as a considerable imbalance between the resources available to (EFL/ESL/ESP) language teachers in terms of teaching materials and those available in terms of testing and evaluation instruments. Furthermore, the observation Jakobovits (1969:63) makes still holds true particularly in testing spoken performance, that is, that developments in the area of assessment have proceeded under the impetus of necessity rather than under the guidance of a coherent theoretical understanding of the nature of language and communication.

This paper describes how UMSEP attempts to redress this imbalance in designing a testing programme and a set of testing procedures that mirror the philosophy behind the approach adopted in the materials that we have developed. It is also our view that a valid testing programme should reflect the course aims and content. Therefore before proceeding to describe the testing procedures in UMSEP, it is pertinent to briefly summarize the main principles underlying the course design.

Rationale for Course Design in UMSEP: A Summary

In our view as presented earlier in 'Rationale, Design and Structure of UMSEP Courses', (this volume) effective performance (the primary aim in UMSEP) does not come with the development of fluency to the exclusion of accuracy, nor with the development of knowledge of formal resources to the exclusion of communicative effectiveness. We therefore adopted an approach with two parallel strands support activities that build up formal resources or competence, and central interactive activities that provide opportunities for appropriate language use or performance. These interactive activities reflect the types of interactions that occur in the target professions.

These parallel strands allow learning to occur through explicit teaching as well as through incidental means, that is, through exposure to language in use, and by focusing the learner's attention on how to use language to solve a problem or to get his meaning across rather than on what particular items to use. The syllabus for each of the three courses was determined through analysis of patterns of interaction and language realisations in the target professions and are specified in terms of functions and interactions.

Having adopted such an approach and having specified content in the manner described above we therefore set about to design a testing programme that would reflect these features. The section below describes the work done to date.

The Testing Programme

In addition to taking into account the materials and methodology we felt that the testing programme should provide the kind of information required by employers in the professional world. Do the employers want the test results to say that the candidate can perform the required tasks, or do they merely want to be told that this person has built up competence and can display this competence in his performance? These two questions have implications for test design and content. As has been pointed out none of the three courses — POSM, OSM and OSL is designed to teach pure performance. The central activities in the OSM and OSL courses approximate target eventsin the professional world. The POSM course meant for final year undergraduates does not need to rigorously follow the demands of the professional world.

With these considerations in mind it was decided that the testing procedures include two main approaches — the Operational Approach and the Discrete Feature Approach. In the Operational Approach the test items built would be facsimiles of real events in the outside world or profession(s). Task types constructed following this approach would be closest to authentic tasks. The Discrete Feature Approach would test communicative knowledge or competence. The tests in this approach would be mainly receptive in nature involving recognition, for example, of forms, functions and other enabling skills.

It was decided that the testing programme for each of the three courses would consist of the following test types

(i) Discrete Feature Tests
(ii) Listening Tests
(iii) Integrated Skills Tests
(iv) Interaction Tests

Each of these types of test is described below

Discrete-Feature Tests

These tests assess the learner's communicative competence. They are mainly receptive in nature. Such tests are given at specific points in the course and reflect closely the language focused upon in the Support activities to ensure content validity They are distinct from discrete-point tests in that they focus on items at the level of functions (and their realisations), and test recognition of forms, functions and enabling skills. The following are two examples:

Example 1

Speaker A. How many visits do you make to the Branch offices in a month?
Speaker B: I visit them twice a month.

<u>Question</u>

Which of the following can Speaker A use to get the same information?

a) How long does it take you to visit the Branch offices?
b) What Branch offices do you visit?
c) How many Branch offices do you have to visit?
d) How often do you visit the Branch offices?

Example 2.

Which of the following could be used to ask Speaker A for more precise information on what he says?

A. I joined the bank as soon as I left the university

a. Did you join the bank when you left the university?
b Which branch of the bank did you join?
c. Where did you work after you left the university?
d. What did you join when you left the university?

Listening Comprehension Tests

These tests involving taped conversations (8-10 minutes) of simulated professional events, reflect the Exposure components of the materials In these tests, students are required to listen and simultaneously answer items that test comprehension of specific information, main points, attitudes, etc. The multiple choice format is often used. Other items test comprehension of arguments in larger chunks of discourse, inference, etc.

Integrated Skills Tests

These tests require the students to demonstrate more than one skill for example, listening and speaking. In such tests the student may be asked to listen to an authentic'or simulated event and to perform specific tasks such as notetaking. They then use these notes to give an oral report. In some tests for the OSL course the law students are asked to take notes in the style that magistrates normally employ The assessment of performance will include the understanding of information in the input and the quality of the spoken language used in the report.

Although the OSL and POSM courses use video as a regular feature in their materials, no tests involving video have been created mainly because of the large numbers of students involved. However, this is a long term objective.

Interaction Tests

The interaction tests which will form the main discussion in the rest of this paper reflect the central interaction activities of the materials.

An interaction activity is defined as one that involves a minimum of two participants working towards an outcome. As in the designing of interaction activities, control of the interaction in such tests is achieved through.

a. providing a clear task to perform (for example, making notes to report a discussion, evaluating alternative proposals);
b providing rules on how to perform the task. Steps and procedures are specified in order to ensure that the right kind of talk emerges and to ensure that each participant has an equal opportunity for talk;
c. providing an information input for example, short texts and/or visuals often as start-off points for talk.

The interaction tests attempt to ensure talk in several ways. Firstly, although the tasks are set in realistic situations which are relevant to the professional contexts, care is taken to ensure that they are not totally unfamiliar to the learners, for example, situations chosen could be a committee meeting to plan a certain event or a research group gathering data. Secondly, concerted efforts are made to avoid unnatural or complicated roles (for example, an irate customs officer) or the need for specialized language. In most cases the tasks allow testees to behave as themselves in hypothetical settings. This is particularly true in the POSM course where the learners have not had any professional experience. In the OSM and OSL courses where some professional 'role-play' is necessary, the testee's ability to 'act the professional role' does not enter into the assessment Thirdly, gaps are created, for example, by requiring testees to take different positions on an issue (an opinion gap) or by giving different information to different testees (an information gap). Often, the testee is required to select or provide information which he 'invents' thus ensuring that the talk is unpredictable.

The interaction tests involve a minimum of two and a maximum of five testees (as it is difficult to monitor more than five) Some tests require the presence of a 'plant' or catalyst who is usually another teacher who participates in the interaction with specific guidelines, for example, to gear the interaction towards a certain course in order to generate the desired type of language use or to introduce an unpredicted element into the interaction to see how the testee copes with the situation.

Interview-Type Interaction Tests

Our interview-type tests differ from the traditional interview situation in which the assessor asks the questions and takes charge of the direction of the discourse. In order to provide the testee with the opportunity to elicit information and demonstrate discourse skills apart from responding to queries, he is given a specific set of instructions which outline his role and the outcome towards which he must work. His interlocutors could be another testee or a plant. This type of test is used particularly in OSL where the testee could be asked, for example, to 'interview' another 'lawyer' (another testee) or 'client' (usually a plant).

Group Interaction Tests

Most of our group interaction tests involve four testees and last approximately 30-40 minutes. We have found that this length of time is adequate to allow an accurate assessment of the testee's performance.

In designing the group interaction tests, care is taken to select a larger 'event' (for example, making and evaluating recommendations) that incorporates interactive activities covered in the teaching units (for example, discussing ideas and opinions, stating a problem and explaining implications, comparing advantages and disadvantages). The following test was used for a block of teaching units on putting forward and evaluating positions in the POSM course:

Situation

The Malaysian government wants to establish a training centre for young men and women. The centre will offer training for skills such as carpentry, television repair, sewing, cooking, and secretarial and book-keeping courses. The centre will be large enough to train 1,000 young people each year However, the government has not decided in which state to build the training centre.

You will be given a card to work with.

PREPARATION (5 mins)

1. Decide: (a) Which state to recommend and which part of the state the training centre should be.
   (b) How to justify your recommendations.
2. Make brief notes so that you can use them for making your recommendation to the group.

INFORMATION

1 Each person will have 5 minutes to present his recommendation.
2. The whole group will have 10 minutes to discuss all the recommendations and to reach an agreement.

Card 1

Your task is to recommend the state of Malacca* for the training centre. You are to get your group to agree with your recommendation.

(*Twelve cards were used, each containing a different Malaysian state)

Assessing Performance

Criteria for Assessment

Several attempts were made following a review of checklists of criteria available (for example, Carroll (1980), Schultz (1977), Jakobovits and Gordon (1974) and the Foreign Service Institute (FSI)) to devise a set of criteria which was felt to be suitable. It was ultimately agreed after trial and error and working with lengthy and detailed specifications that performance could be effectively assessed on four criteria, that is, fluency, effectiveness, accuracy and range, defined as follows.

1 Fluency· smoothness of delivery without obvious groping for words, continuous and natural speech, does not refer to speed of delivery
2. Effectiveness. ability to convey meaning and make oneself understood

without causing stress on the listener; ability to employ appropriate turn of phrase and precise vocabulary

3 Accuracy: ability to produce grammatically correct utterances; pronunciation does not prevent understanding of what is said by any good speaker of the language: ability to interpret correctly structural and lexical range.

4 Range: the ability to draw upon a wide range of discourse skills, structures, as well as lexical items.

It can be seen then that aspects of both competence and performance are taken into account in the assessment. It will be noticed that each criterion subsumes other performance factors For example, fluency subsumes 'repetition' and 'hesitation' which are listed as separate criteria by Carroll (1980) and effectiveness incorporates 'flexibility', 'independence' (Carroll, 1980) as well as 'amount and quality of information' (Shultz, 1977). This is evident from a look at the Rating Scale below

The Rating Scale

A six-band Rating Scale is used in assessing performance. Two of the bands, that is, Band 3 (the Criterion level) and Band 6 (the Target level) for the POSM course are given in the two tables on page 67

The Criterion level, Band 3, is established as the minimum level of acceptable performance in terms of course aims and professional target settings Band 6 is the Target level which, as can be seen from the table, is not equivalent to 'native' or 'near native' levels. This is felt to be a realistic objective givea students' proficiency levels at entry and constraints within the local situation.

Testees are placed on the Rating Scale as follows.

Band	6	Good Speaker
	5 ·	Competent Speaker
	4	Modest Speaker
	3 ·	Marginal Speaker (Criterion or 'Pass' level)
	2	Extremely Limited Speaker
	1	Intermittent Speaker

Apart from the Rating Scale 1, overall performance descriptions for each of the six bands have been drawn up. These descriptions specify in greater detail the characteristics of performance at each level and are used basically as a source of reference particularly during the training of assessors Overall performance descriptions for Bands 3 and 6 for POSM are as follows.

Band 3 Marginal Speaker

Speech is maintained at a continuous though uneven pace. Is sometimes hesitant and unsuccessful at finding the right words Has enough mastery of basic structural patterns and knowledge of a few complex ones to give the impression of being able to cope without great difficulty

Vocabulary is broad enough for general needs and includes a basic working knowledge of professional lexis. Errors in use of basic stress and intonation patterns and pronunciation but this does not cause serious misunderstanding, Band 3 and Band 6 [POSM]

Band	FLUENCY	ACCURACY	RANGE	EFFECTIVENESS
3 CRITERION (PASS)	1 Able to keep communication going although speech is uneven, hesitant and marked by some unsuccessful groping for words. 2. Some unnatural pauses and false starts. 3. Able to rephrase ideas although with obvious effort. Some utterances may be left unfinished.	1 Basic structures: shows ability to use most of them but errors are made. Complex structures: able to use a few but problems remain. 2. Vocabulary: shows mastery of basic vocabulary. Some incorrect use of basic professional lexis. 3. Uses most stress and intonation patterns accurately. Pronunciation does not cause serious misunderstanding.	1 Depends mainly on basic structures. Attempts some complex structures with limited success. 2. Vocabulary: adequate for personal and daily needs. Adequate working knowledge of basic professional lexis. 3. Uses basic stress and intonation patterns to make meaning clearer 4. Does not always use short utterances. Has some basic skills for interaction.	1. Is usually able to achieve his goals but with some noticeable difficulty. 2. Takes some initiative to make clear difficulties and to clarify meaning. Relies on interlocutor in many cases. 3. Shows some ability to cope with the unexpected.
6 TARGET	1. Speech is smooth with the occasional hesitation and slight groping for words. 2. No stumbling or unnatural pauses. 3. Able to maintain communication without difficulty.	1. Basic structures: good command. Complex structures: uses most accurately 2. Vocabulary: good command of basic lexis and basic professional words. Accurate use of much specialised lexis. 3. Makes occasional errors in stress and intonation. No noticeable problems with pronunciation.	1 Uses most basic structures with ease. Demonstrates ability to use many complex structures. 2. Vocabulary: wide range of general and most professional lexis. 3. Able to vary stress and intonation to convey meaning effectively. 4. Has adequate skills for most interactions.	1. Able to convey information in most cases without difficulty. Achieves his goals with little effort. 2. Is mainly independent of the interlocutor Able to initiate and direct interaction to desired goals. 3. Is able to handle the unexpected with success. Occasional problems with unfamiliar vocabulary and concepts.

Able to understand most parts of normal speech on non-technical subjects. Needs interlocutor to repeat or clarify especially when discussing topics of a less general nature.

Can handle with some confidence most familiar topics. Is flexible enough to cope with the unpredictable but with limited success.

Is able to seek, convey and evaluate basic information effectively and can put forward and evaluate positions. Is able to react to proposals in order to reach a decision but has only limited language skills for negotiation.

Band 6: Good Speaker
Speech is smooth, marked only by the occasional hesitation and slight groping for words Able to use language fluently and accurately at all levels pertinent to most practical and general professional needs Breadth of vocabulary covers most professional lexis Errors of grammar and pronunciation are minor. Is effective in varying stress and intonation to convey meaning.

Can comprehend and keep pace with any interaction at all levels. Maintains independence consistently Able to effectively direct interaction to achieve goals.

Maintains high level of participation on all levels even in unfamiliar situations Is sensitive to and able to cope with attitudinal tones and unpredictable changes in topics. Maintains inputs at a clear and logical level and can direct interaction coherently and constructively

Can initiate and develop topics in information-sharing interactions without difficulty. Is able to support his position effectively and evaluate other positions appropriately Shows a command of language skills required for straightforward negotiations.

The above procedures and test types were devised after much trial and error, rethinking and revision. The following section briefly outlines the initial test construction, piloting and measures that were taken towards developing the final testing package described in the preceding section.

Initial Test Construction and Piloting

Although the following section describes earlier developments and not the current state of the UMSEP testing programme, it is included here so that we may share with others who may be involved in similar work in the field, the types of concerns and problems we encountered in the earlier stages of test development. In early 1982 initial tests were constructed on the basis of the specifications outlined in earlier sections of this paper These tests were administered at the end of the pilot courses in April and August 1982. This section will focus on aspects of student performance on these tests and questions related to the validity and reliability of the testing procedures.

Piloting — April 1982

I The performance of students on the Oral Interaction tests was as follows. (as measured on the six-point scale)

POSM	Mean score = 3.2
OSM	Mean score = 3.1
OSL	Mean score = 3.5

The majority of the candidates in the POSM course obtained a score of 3 which is a Criterion level score. There seemed to be a bunching of scores on the 2nd, 3rd and 4th points on the six-point scale. This was largely due to the low entry levels of the students although there is the possibility that there could have been some reluctance on the assessors' part to award very low or very high scores.

2. In the discussions that followed a number of issues related to the validity and reliability of the test instruments were raised. The first of these touched on the authenticity of the tasks set in the listening comprehension tests for the OSM and OSL courses and the extent to which these could be classified as communicative.

The main criticism levelled at the Oral Interaction tests was that the notes in the student briefs were too technical in nature. There was too much of input in terms of the information provided. It was pointed out that a student's performance on the test depended very much on whether he understood the notes given in the brief The original intent was that the notes in the brief should aid the student in his performance In practice it had turned out to be otherwise. Here was a fault in the test design which needed correction. In the light of this feedback it was decided that subsequent group oral interaction tests would keep the input in student briefs to a minimum The topic chosen for the interaction should also not be too demanding in terms of lexis

The question as to how dependable the scores on student performance were, raised the issue of the relialibility of the test instruments. It was felt that the multiple choice items in the listening and discrete feature tests ought to be pre-tested and item analysis carried out to ensure item quality and improved reliability

The key to achieving objectivity in the scoring of the Oral Interaction tests rested on the efficient use of the checklist of critical performance factors that were to be observed in the oral interaction and the rating scale used to arrive at a global rating of the student's performance. This would depend largely on the consistency of the rater's or assessor's judgements The briefing sessions called to familiarise assessors with the checklist of performance criteria and the six-point rating scale though adequate, needed to be improved upon

Another point raised was whether the plants were sufficiently aware of their role in the test and whether the training provided was adequate to minimise 'inter-plant' and 'intra-plant' variance in eliciting the desired performance from the students.

Administrative and procedural details like the duration of tests and the quality of hardware and software used in tests were matters that needed to be looked into.

Piloting — August 1982

In the light of experience and feedback from the previous piloting session measures were taken to ensure a more valid, reliable and efficient testing programme.

The following measures were taken to ensure test validity The course writers were asked to study the tests to verify the relevance of these to course content. The tasks set were selected on the basis that they represented professional needs or events in real life. Expert opinion was also sought on the authenticity of these tasks

In the case of the discrete feature test and the listening test, item analysis was carried out for the multiple choice items. These were reviewed in the light of item statistics and those found to be wanting were discarded. A system for the conversion of the six-point scale to raw scores to fit requirements of various faculties was worked out. A standard marking scheme was also drawn up for assessing the integrated skills test. Further measures to enhance the reliability of the testing programme involved the training of assessors and plants.

The Training of Plants

The training of plants or catalysts involved first of all a definition of the role. The plant or catalyst would be someone familiar to the students being tested, preferably the teacher One of his main functions is to put the testee at ease and provide help when it is required. He assumes the role assigned to him in the interaction and gives the testee opportunities to display his competence. He could accomplish this either by asking probing questions, or by being purposefully evasive in withholding information. He is also seen as the instrument for introducing unpredictability into the interaction to see how the testee copes with the situation.

Before the tests were trialled 'would be' plants — teachers who had taught the pilot courses — were exposed to recording of oral interaction tests. They were asked to evaluate the performance of the plant in the test. It was found that some plants tended to dominate the interaction while others provided the testees with the desired opportunities for performance. In this manner plants were made aware of their role and function. They then simulated the role of the plant in a group oral interaction test.

The Training of Assessors

Before the Oral Interaction tests were administered assessors were brought together They were provided with the six-point rating scale and descriptions of the criteria on which the oral interactions were to be assessed — fluency, accuracy, range and effectiveness. They went through the Criterion and Target level descriptions of the criteria and doubts were ironed out. Assessors were told that they were required to make a global assessment and not to award separate scores for each criterion.

Next, the assessors listened to recordings of student performance in a group oral interaction test (April 1982 Piloting). Assessors were asked to rate the performance of one selected student bearing in mind the Criterion and Target level descriptions of the criteria. The scores awarded for this candidate were as follows. 3, 3, 3, 3, 3, 5 The majority of the assessors had awarded this candidate a score of 3 - a Criterion level score on the six-point scale. In the discussion that followed it became apparent that the assessor who awarded a score of 5 recognised the student's voice and awarded the score on the basis of his performance in class. This assessor agreed that she had been overgenerous in her assessment

The scores awarded for the second candidate who was assessed were as follows. 4, 4, 4, 4, 4, 3. There was general agreement among the assessors that this candidate's performance was above that of Criterion level and that he deserved a score of 4 on the six-point scale. From the trial grading procedure it was evident that assessors had a fairly good grasp of what is expected at Criterion level performance on the four criteria.

Overview of Pilot Testing Programme — POSM, OSM, OSL

There is a need for continued activity in the testing area for the three courses More tests need to be written and pretested on larger samples of the target population. There is also a felt need for a more rigorous item analysis of the tests piloted. This is essential in terms of establishing the validity and reliability of the measurement tools used. This activity would lead naturally to the building up of an item bank for each of the courses — a bank of items of proven quality

It is also necessary to think further about the training programme for assessors and plants for the three courses It is noteworthy that the first step in this direction has been taken — the selection of samples of Criterion and Target level performance. Logically, the next step should be the institution of an on-going programme for training.

Summary

This paper has described briefly the testing procedures used in the University of Malaya Spoken English Project. It has shown that the testing procedures attempt to reflect course goals and objectives, and incorporate the assessment of both competence and performance. Various tests have been outlined while the main focus of attention has been on the Interaction Tests

It was pointed out that the types of interactions in the materials and tests are modelled on target professional events. Performance is assessed on four criteria specified in the six-band Rating Scale. The specifications of criteria and Rating Scale in use currently have been refined through trial and error This represents the attempt at improving reliability which has also included the training of the assessors and the plants.

The tasks ahead for the UMSEP team in the area of testing are many. We are, for example, in the process of organising a bank of items particularly on the discrete-feature tests Another area of work is correlating our tests to standardised proficiency tests. Although there are many related issues as yet not completely resolved, we have found that testing performance of a large population in our situation through the procedures described has provided us with a more than fair means for accurate assessment.