The Problems of Oral Testing. What Did you Say?

Mike Nicholls

The Problems of Oral Testing. What Did you Say?

Background

Whilst preparing this talk, I came across a quotation that seems to me to sum up the problems involved in oral testing rather nicely. It comes from a paper on the classification of oral competence published in 1981 by two Americans, Madden and Jones:

> "During the past few decades, oral language testing has had a great deal in common with physical fitness. Everyone thinks it is a wonderful idea but few people have taken the time to do anything about it. "

I am afraid, twelve years on, that is still largely true. There has been a lot of research done in the past decade in the area of oral competence testing. But nearly all of it has been done in the States and has suffered from what I tend to regard as the American research disease, the overriding concern with statistical information, the need to make things statistically sound without necessarily having very much to do with what goes on in the language teaching classroom or in the language acquisition of students. We have learnt a lot about how we can use factor analysis, predictive validity and constructs of this nature. But we have not done a great deal until the last two or three years in looking at what needs to be done to create oral language tests, to actually measure the language competence of our students in doing the things that they want to do in the language.

However, in the last four or five years, the University of Cambridge Examination Syndicate has at last begun to get involved in this area. I say, "at last", because Cambridge has been examining oral English since 1913 on a world-wide scale. And yet, it was only in 1989 that they were finally persuaded to establish an English Language Testing Research Division, when the testing research department was set up in the EFL (English as a Foreign Language) division, headed by Michael Milanovic. Since then it has grown to an eight-man department. It is spending a lot of time and money on research and it has set up a number of very interesting research projects. In particular it has become concerned with what happens in oral language testing and they have recently proposed a new model of oral testing.

I now propose to take you through this model and to look at the constraints that affect language testing, and the factors that need to be taken into account. I will then look briefly at the latest English language test that has

come into effect in Cambridge, the Certificate in Advanced English (CAE) and at how oral language competence is assessed in that exam.

A rational model of the test development process (*Chart 1*)

The process starts with a *requirement for a test*. Unfortunately, though it appears self-evident that this should be the prerequisite factor, decision makers do not always start with an actual decision that a test is needed, not that it would be nice to have a test, but that there is a need for a test.

When the need has been established, it is succeeded by the *planning phase* in which a situational analysis is carried out and a project plan is written - in which a time scale is included.

Following on from the project plan comes the *design phase* in which the initial test specifications are drafted. This process should then be reconsidered (going back to the planning stage, reviewing the considerations and constraints, and evaluating the test design and content specifications) before the production of sample materials commences.

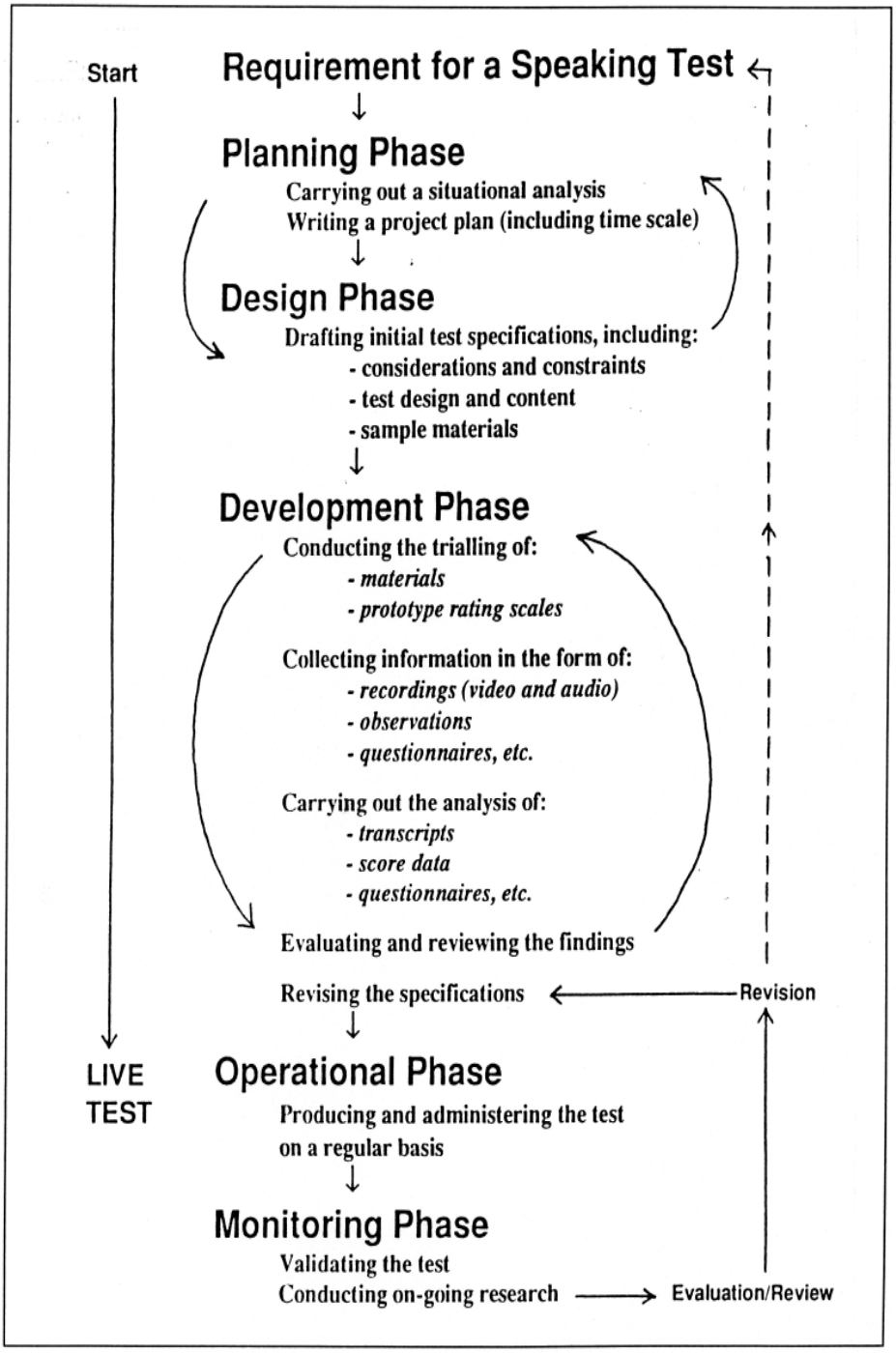
After the sample materials have been prepared, the *development phase* begins in which the trailing of the sample materials and of the prototype rating scales is undertaken. Information is collected on the efficacy of the test and its procedures, usually in the form of audio and video recordings of what goes on in the oral interview, and of the observations of the oral interviews made by outside observers, by the interlocutors and assessors and by the interviewees. Finally, questionnaires are handed out to everybody concerned at each level, to obtain as full an appreciation of the test as possible, covering all aspects of design, administration and operation. The transcripts, the score data, the questionnaires and observations, are subjected to analysis to evaluate and review the process. Dependent upon the outcome of these evaluations, the design phase may need to be repeated and aspects of the test may require redesign, redevelopment and re-testing until finally a set of specifications is achieved which satisfies the original requirement and is consistent with the constraints and considerations operating upon the administration.

This process being completed, the *operational phase* of the test in which the test is produced and administered on a regular basis can safely be entered into. There is necessarily a continuing *monitoring phase*, which is concerned with validating the test, conducting ongoing research with regard to it, and a process of evaluation and review which leads to regular revision.

Chart 1

ASSESSING SPOKEN LANGUAGE

A Rational Model of the Test Development Process



Michael Milanovic and Nick Saville UCLES EFL Division October 1993

This is the model currently used in Cambridge, designed for tests taken each year by 350 000 candidates now. However, it is also a model which can be used just as effectively on a much smaller scale within a small institution. It is feasible to do it at any level and it is necessary, if you are going to have a test of any value, that this process is carried out conscientiously in the design and development of any new oral test.

Let us now look at the elements of this model in more detail. The planning phase opens with the situational analysis.

The situational analysis

What information is needed in order to carry out a situational analysis? The main activity is the identification of the constraints operating upon the situation.

It is necessary to identify, first of all, the *stakeholders*, the people who have an interest in the form of test which results from the process, *i.e.* those involved in the testing process, including not only the candidates but also the teachers, the management of the institution, the parents, the employers, the government agencies, etc. It is principally a question of who takes it, why they take it and what information everybody gets from it. For all of these interested parties, one needs to get some measure of acceptability of the test that is being produced, what it is they want to actually have.

Secondly, the *purpose of the test* must be considered, that is the reason for developing it, the way the test should fit into the current system (in terms of curriculum objectives, current practices, future directions) and the level of difficulty for the intended test takers.

The third constraint is to look at the *extrinsic factors*, the factors outside the institutions:

- external expectations of how speaking should be tested (*e.g.* the commercial testing market, the availability of other tests of speaking, local testing experts, etc.)
- societal demands (*e.g.* the socio-economic climate, educational policy, local conditions, etc.)

I remember when on my first official visit to France on examination business in 1988, although the specifications for the CAE insisted that all interviews should be carried out in pairs, unanimous feedback was received

indicating that pair interviews would not be acceptable. It was said that the expectations of an oral interview test were that there would be one interviewer and one candidate and it would not be acceptable to introduce a two interviewer/ two candidate examination. Interestingly, France was one of a very small minority of countries amongst the many consulted which held this view. Another one was Japan. I have noted a number of similarities between the systems in France and Japan which has surprised me. The rigidity of the system and the expectations of everybody outside of the process is so much greater in France than almost everywhere else. In most countries, only teachers and candidates are greatly involved in worrying about the exams, and perhaps to a minor degree the school administration . But in France, parents care, employers care, the man in the street has strong views about what happens in language examinations. It is amazing the amount of heat that can be generated by discussing language examinations.

Finally, we have to examine the *intrinsic factors*, the institutional parameters within the school or college.

- The existing working practices of the teaching staff.
- The knowledge of theories related to the testing of speaking amongst, not only the teachers preparing the test, but all their colleagues.
- The availability of resources (human, technological, temporal and financial) for the development of the test, its administration, the reporting of the results to test takers and users, the replication of the test, because you cannot function with just one test version (if you want to carry out a long-term testing programme you need a number of test versions that are of the same level of difficulty) and validation of the results.

Facets which interact in the testing of speaking

Having completed the situational analysis, the next matters requiring consideration are the various facets which interact in the testing of speaking.

- *The candidates* bring to the test their own background, and that can be a diffi- culty when examining. Examiners find themselves faced with candidates with a wide range of ages and backgrounds, a candidate in his sixties, with forty years of life experience as an adult, has a lot more to say about most issues than an eighteen-year-old. The behaviour of candidates, within an examination context, is also important to take into account. The samples of language that are

elicited are very important. The essential concern of the examiner and of the test is : what do the candidates say and how can they be stimulated to produce language which is appropriate and suitable for marking and grading?

- *The examiners* raise similar problems of age, background and examining style There is a very wide variety of styles of interviewing, and some examiners do very much better at getting the sort of responses that the exam expects of the candidates than others. Some people are very good at sitting and encouraging, primarily by body language : the raising of an eyebrow, the curling of a lip, etc. Others seem to need to talk immensely and if you have only got fifteen minutes interview, you do not get much language out of the candidate, if you do talk at great length. Another problem is the sort of language examiners use. For the CAE, Cambridge now produce a script for the first part of the exam which says : "Introduce yourself. Say 'Hello, I am... What is your name?' etc." to try and standardise the language used by examiners, the verbal prompts that are used, and how much spontaneous talk the examiner allows himself. It is difficult to standardise and control; a good, intelligent candidate, with relatively limited English and a great awareness of his weaknesses, will quickly try and find ways in which he can encourage the examiner to talk at length and you would be surprised how good some of them are at making examiners talk. The rating scales are another problem. What is it that is actually being tested? A very clear picture of what it is that is being measured is necessary.
- *The tasks* that are set are another important variable. Should the test utilise principally verbal prompts of a fixed or free format or are visual prompts, written or pictorial, preferable? The range of task types which can be utilised in oral language testing are numerous. Underhill (1987) gives a list of twenty-two different task types with variations of about sixty-six different possible ways of eliciting within an examination context. These are discussed below.
- *The ratings* are a further variable of some complexity. The nature of scales, their number and length offer a wide range of choice to the test constructor. Some examinations use a single holistic scale whilst others use different scales for the different skills being evaluated. In the United States, the Foreign Language Service Test uses a single holistic scale. Cambridge have always used a set of five or six individual scales. In FCE (First Certificate of English) and CPE (Cambridge Proficiency Examination) interviews, lasting for ten to twelve minutes, the single examiner has to measure on five or six different scales of five or six points each as well as manage the interview in such a way as to stimulate appropriate samples of language for assessment. The Cambridge Proficiency is particularly difficult. There is a maximum of fifteen minutes with

one examiner and one candidate. There are six different scales to apply and the candidate usually produces a considerable range of language. At the same time, the examiner has both to concentrate on what he is saying, on directing the discourse and on assessing it. At the lower, FCE, level, this is not as difficult. The amount of language produced and the variations in it are more limited and therefore the examiner's task is less Herculean.

With the recently introduced CAE examination Cambridge have combined the holistic and individual skills scales approaches by installing five individual scales plus an overall holistic scale. There are two candidates and two examiners, one of whom acts as the interlocutor, doing all the talking and questioning and bringing out the language. The second examiner acts as the assessor who concentrates solely upon the assessment of the language produced by the candidate. The interlocutor assesses the candidate but only on the global scale, giving a single overall mark, the assessor marks on each of the six individual scales as well as giving a global mark. This method works very much better in terms of getting consistent marking and leaving the examiner feeling contented that he has been able to concentrate on the task of satisfactorily examining the candidates.

Checklist for the development of speaking tests

The Cambridge model, as presented by Milanovic and Saville (1993), also offers a checklist for the development of speaking tests. There are two sections to this checklist

Professional considerations

- *Target language use :* In what kinds of situation do candidates need spoken English? If there is a small number of variations in language, it is possible to work out the target language required very quickly. However, with an examination that is supposed to cover overall language proficiency at an advanced level, the problem is much more complex. The great difference between the Cambridge Proficiency exam and the other Cambridge exams is the much wider range of abstract thought and expression of abstract ideas required by the examination. Therein lies the difficulty in designing the materials for the test.
- *The level of spoken language performance* which is necessary for these situations needs to be assessed in terms of range, accuracy, fluency and appropriacy.

- *Real-world language events* need to be recreated in the testing context. They need to be defined in terms of their physical settings, the channel to be used (face-to-face, telephone, etc.), the real time processing requirements involved, the interaction pattern concerned (degree of participation), the number and type of interlocutors involved (status, gender, age, familiarity), the purpose of the event, the topics which are to be covered, the tasks which need to be performed and the amount of language to be elicited.
- *The information to be given to test users*. It is necessary to decide how much and what type of information on performance is going to be fed back to the users of the test. The possible approaches include a wide range of options from simply the release of pass/fail information through the release of total scores, the release of scores for each part of the test through profiles based on task achievement or statements of ability displayed by skill or test section.

One of the most common complaints that the Cambridge Examinations Syndicate receives from test users is that not enough information is provided to them on their performance in the tests. The decision has been taken in Cambridge to give the candidate a much fuller profile of his results. That cannot come into effect before 1995, simply because the mechanism of recording the information and producing it has taken a six-year development period to achieve. When information concerning the results of 400 000 candidates a year has to be produced, it is important that it should be helpful and correct.

Practical considerations

Administration

The structure and content of the tests being designed must also reflect the practical constraints applying to the administration of the test. These include such factors as manpower and premises.

- Are there sufficient qualified and experienced professional staff to design and implement the test?
- Are sufficient staff available to help conduct the tests?
- How many rooms are available for conducting speaking tests?
- Over what period must the tests be completed?
- How long can each test session be?
- How quickly do the results of the test have to be issued?

All these are essential constraints on testing within any institution and nearly all of them reduce what can be done and quickly reduce ambitious plans for oral testing to what can actually be achieved in the time available.

Candidates

Another set of constraints applying to the test development process is that concerned with the candidature, not the skills and learning brought to the examination room by the candidates but the physical constraints affecting the exam which arise from the structure of the candidature.

- How many candidates need to be assessed?
- How long should the assessment be for each candidate?
- How are candidates to be interviewed? Individually? In pairs? In groups of three or more? If in pairs or groups, how will the pairings/groupings be made? Paired and grouped interviews for the FCE and PET (Preliminary English Test) have been introduced in Toulouse in the past year, as an option, and students are allowed to opt for being interviewed in pairs. They are asked if they have a friend with whom they have been studying and with whom they would like to be interviewed. They are being encouraged to opt for the pair interview format because it is believed to reduce stress and because Cambridge have decided to gradually convert all their oral examinations to a two-examiner/two-candidate format over the next few years in the interests of increased consistency. More and more local candidates are taking up this option voluntarily.

Examiners and ratings

A further set of practical constraints are concerned with the availability and quality of oral examiners and their skill in applying the rating scales being utilised. The considerations here include :

- How many examiners are available?
- How many candidates must be assessed per hour per examiner?
- How many examiners should be there for each assessment? One examiner? Two examiners? More than two examiners?
- Is it possible for all examiners to be native speakers?
- If not, is it possible to find examiners who speak the language well?
- How are the examiners to be trained? How much time is available to train them? Those two constraints are always in conflict. Ideally, with any new examiner a minimum of at least one day, full time, for going through the classroom training process is required. At least another half a day of that examiner's time, acting as an observer, and then being observed examining in real live examination situations is also desirable. It is often difficult to get a day and a half of training time for the number of examiners required. However it is

very difficult to attain satisfactory standards of examiner performance exams unless examiners have been adequately trained for the task.

- If insufficient examiners are available, is it possible to use taped input? This is often seen as a very attractive possibility, but a problem with taped input is that the tasks are necessarily of dubious authenticity, in that there is very little, in terms of conversational use of English, that does not vary with the response given. So, whilst discrete item type testing on tape is possible, any realistic conversational development is impossible. There is therefore much that cannot be satisfactorily tested if taped input is used. If the other constraints dictate its use, it is necessary to reduce the range of language to be assessed, and therefore to set targets and tasks susceptible to stimulation by taped input material.
- If you are going to do tape-based tests, is it possible to use a language laboratory ?

Tasks and materials

The selection of the materials to be used and the tasks to be performed has already been raised above. In deciding on these, the following questions have to be answered.

- How many *phases* will be used in the assessment? The average Cambridge exam uses three or four phases. In the exam of the FCE, there is a photograph to talk about, followed by passages to be read silently and then assigned to different photographs, followed by the simulation task or discussion.
- Will *interlocutors frames* be used to guide the examiners? Are the scripts tight or fairly vague and used by examiners to keep their language within tolerable limits?
- What sort of *tasks* will be used in each phase?
- Will photographs and other graphic prompts be used?
- Will other kinds of prompt material be used?
- How will equivalent sets of materials be produced?

The range of task types which can be used have been mentioned above but include:

- Discussion/conversation between examiner and learner.
- Discussion between learner and learner.
- Oral reports.
- Discussion and decision-making.
- Role plays.
- Straight interviews.

- Description and re-creation. An example of this kind of task is that used in the CAE when one candidate is given a picture, a simple line drawing of some kind, and told to describe it to the other candidate who has a piece of paper and a pencil. The idea is that the second candidate can draw to the instructions of the first and will end up with something that looks like the piece of paper provided to the first candidate. A difficult task, but interesting for stimulating instrumental language.
- Appropriate responses, these tasks are appropriate to taped input.
- Question and answer.
- Reading blank dialogues.
- Using a picture or picture story and creating a story from it, giving details of the pictures seen.
- Giving instructions.
- Précis or re-telling of a story from an oral prompt.
- Reading aloud. It is not a necessary skill for many people, but it is a way of getting a suitable sample of language, for assessing phonetic features in particular.
- Translating and interpreting at high levels. Candidates translate from a written text into oral English or listen to a tape or to something being said by the examiner and interpret almost simultaneously.
- Sentence completion tasks.
- Sentence correction.
- Sentence transformation.
- Grammatical exercises

and many, many more.

Assessment

The administrative elements of the assessment process constitute a further set of practical considerations which must be considered before a test is constructed. They include the following items:

- How will the scores be reported to users?
- How many ratings will be made for each candidate?
- Will discrete-point assessment be used?
- Will holistic assessments be made? If yes, will a single overall ability scale be used?
- Will component scales be used? Are there several different scales for several different skills? If so, how many scales will be used?
- What are the scales to be called?

- Is partial credit scoring used?
- How will the reliability of scoring be ensured?
- Is it possible to make recordings for second or third ratings?
- Who will make the additional ratings from tape? Candidates are often put off by taping if they are aware of it and they are upset if they discover later that they were taped unawares.

Quality control procedures

There are further considerations that need to be borne in mind if it is decided to make recordings of assessment interviews for checking the consistency of assessment.

- Is it possible to make recordings for quality control checking?
- Who will check the tapes?
- What other methods can be used for quality control checking?
- How will this data be collected and stored for analysis and validation?
- Who will carry out the analysis and validation?

Facets of performance testing : interrelation (*Chart 2*)

The chart showing the interrelationships involved in the facets performance testing is intended to demonstrate the range of features which need to interrelate in the process shown in the model.

The ***examination developer*** produces a *specifications construct*, a design for the test which results from the situational analysis and the physical constraints and practical considerations. This construct specifies:

- *the examination conditions* which will apply when the candidate takes the examination;
- *the tasks* to be used in the assessment interview which utilise the materials provided by the examination developer in an event which involves exchanges of language between the candidate or candidates and the interlocutor/examiner;
- *the assessment criteria* to be used by the assessor/examiner;
- *the assessment conditions and the assessor training* which result from the professional and practical considerations implicit in the test design.

The ***candidate*** brings to the exam his *knowledge and ability*. This is then applied to the task set for him/her by the developer and applied to him/her by the examiner(s). The *sample of language* which the candidate produces as a result of this interaction is subjected to the *assessment criteria* and the ***examiner***, with

Chart 2

FACETS which interact in the testing of Speaking

Candidates

- background
- behaviour
- sample of language elicited

Examiners

- background
- behaviour
- language
  - instructions
  - verbal prompts
  - spontaneous talk
- ratings

Tasks

- verbal prompts
  - fixed format
  - free format
- visual prompts
  - written
  - pictorial

Ratings

- nature of scales
- number of scales
- length of scales

Michael Milanovic and Nick Saville UCLES EFL Division October 1993

due regard to the *assessment conditions* and applying the *training* he has been subjected to and the *knowledge and ability* which s/he brings to the assessment, produces the *score*. All these elements are interrelated and it is their interaction which results in the desired outcome, the satisfactory and consistently reliable assessment which was the aim of the test developer at the outset of his task.

Reliability and validity

In terms of validity, that is whether the exam measures what the people use the language for, the great argument that has raged across the Atlantic over the last decade about examining has been the question of reliability and validity.

- *Reliability* is concerned with whether the results of a test are replicable, *i.e.* whether if the test is given on a number of occasions to the same candidates the results will remain constant. *Validity* is concerned with whether the test measures features which are directly relevant to the testee's ability to perform appropriately in the language in real world situations. American researchers have tended to agree that an exam must be reliable above everything and the Cambridge exams are often considered by American researchers to be inherently unsatisfactory because they do not place primary emphasis on reliability (L.F. Bachman).

In general, most British researchers have agreed with Cambridge that, whilst reliability is important, validity is more so. They appear to consider that discovering what candidates can do with the language in authentic situations is the primary goal of oral language testing and that therefore the first concern of the language tester should be the validity of the testing exercise. There are various forms of validity which have been considered in discussing the value of various forms of oral assessment.

- *Face validity* can be simply defined as: do the students think that the test is actually useful? Does it test what they expected it to test? Does it look the sort of test they think they should be taking?
- *Content validity* is basically: do the tests test what the experts think the test is testing? Are they relevant?
- *Construct validity* is concerned with whether the test is consistent with an underlying theory. These are all subjective measures.
- *Concurrent validity* is an attempt to provide a more objective statistical approach by comparing the results of a test with a previously administered test of a different type intended to evaluate according to the same criteria.

- *Predictive validity* attempts to measure whether the test actually indicates how well the candidate will perform in real situations.

Cambridge has funded considerable research in the past few years in an attempt to find some way to reconcile these views. There has been considerable study of the content of examinations and what examination results actually show. As a consequence, there are signs that the great methodological divide on this issue may be gradually drawing together. The ideal test would be and would be seen to be a good measure of the ability to perform in the target language in the real world and would also provide a consistent and replicatable score.

Conclusion

I think oral testing is like physical fitness in being generally lauded in principle and generally ignored in practice. If we talk about the language in layman's terms, oral testing tests what people do with the language. The testing that we do most of the time is written testing, which is virtually irrelevant to most of what people normally do. Most of our students, when they leave their institutions, will not use the language in its written form or very rarely. They will probably not do a great deal of reading but if they are using the language at all they will have to listen a lot. Yet, at least 80% of what teachers do in assessment terms is based on the written language and not on the spoken language. Oral language is more important in everyday use than the written language for L2 (second language) learners as for L1 (native language) learners. We tend to spend far too little time looking at how oral competence is assessed because it is much easier to prepare written tests than oral ones. A major reason for this may be that written language tests do not have to be done in real time. They are much easier to control, easier to set rules for, etc. But are they helpful or less helpful to the student? Oral testing is going to become more important during the next decades and it is going to be difficult to get oral testing right.

ANNEXES

Appendix 1: **References and bibliography**

Bachman, L.F., and Palmer, A.S., 1981, *A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading*, in Palmer, A.S., Groot, P.J.M., and Trosper, G.A., eds.

Bachman, L.F., and Palmer, A.S., 1983, *The construct validity of the FSI Oral Interview*, in Oller, J. W., ed.

Bachman, L.F., 1990, *Fundamental considerations in language testing*, OUP, Oxford.

Beebe, L.M., and Zuengler, J., 1983, *Accommodation theory: an explanation of style shifting in second language dialects*, In Wolfson and Judd, eds.

Beebe, L.M. and Takahashi, T., 1989, *Do you have a bag? Social status and patterned variation in second language acquisition*, In Gass, S, Madden, C, Preston, D., and Seliker, L., eds.

Berwick, R. and Ross, S., 1993, *Cross-cultural pragmatics in oral proficiency interview strategies*, Paper presented at the 15th annual Language Testing Research Colloquium, Cambridge.

Clarke, J.L.D., 1988, *Validation of a tape-mediated ACTLF/ILR scale based test of Chinese speaking proficiency*, Language Testing, 5.

Dandonoli, P., and Henning, G., 1990, *An investigation of the construct validity of the ACTFL oral proficiency guidelines and oral interview procedure*, Foreign Language Annals 23.

Engelskichen, A., Cottrell, E., and Oller, J.W., 1981, *A study of the reliability and validity of the Ilyin oral interview*, in Palmer, A.S., Groot, P.J.M., and Trosper. G.A., eds.

Gass, S, Madden, C, Preston, D., and Seliker, L., eds., 1989, *Variation in second language acquisition volume I: Discourse and pragmatics*, Multilingual Matters, Cleveland, Avon.

Jones, E.E. and Gerard, H.B., 1967, *Foundations of social psychology*, Wiley, New York.

Lazaraton, A., 1993, *A qualitative approach to monitoring examiner conduct in the Cambridge Assessment of Spoken English (CASE)*, Paper presented at the 15th annual Language Testing Research Colloquium, Cambridge.

Linacre, J.M., 1989, *Multi-faceted Rasch measurement*, Mesa Press, Chicago.

Lowe, P., 1981, *Structure of the oral interview and content validity*, in Palmer, A.S., Groot, P.J.M., and Trosper. G.A., eds.

Lumley, T.J.N. and McNamara, T.F., 1993, *Rater characteristics and rater bias: implications for training*, Paper presented at the 15th annual Language Testing Research Colloquium, Cambridge.

McNamara, T.F., and Lumley, T.J.N., 1993, *The effect of interlocutor and assessment mode variables in offshore assessments of speaking skills in occupational settings*, Paper presented at the 15th annual Language Testing Research Colloquium, Cambridge.

Magnan, S.S., 1987, *Rater reliability in the ACTFL oral proficiency interview*, Canadian Modern Language Review, 43.

Milanovic, M., Saville, N., Pollitt, A and Cook, A., *Developing Rating Scales for CASE: Theoretical Concerns and Analyses*, Paper presented at the 14th annual Language Testing Research Colloquium, Vancouver.

Oller, J.W. ,ed., 1983, *Issues in Language Testing Research*, Newbury House, Rowley, Massachusetts.

Palmer, A.S., Groot, P.J.M., and Trosper. G.A., eds., 1981, *The Construct Validation of Tests of Communicative Competence*, TESOL, Washington D.C.

Ross. S., 1992, *Accommodative questions in oral proficiency interviews*, Language Testing, 9.

Ross. S. and Berwick, R., 1992, *The discourse of accommodation in oral proficiency examinations*, Studies in Second Language Acquisition, 14.

Schmidt. R.W., 1993, *Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult*, in Wolfson and Judd, eds.

Shohamy, E., 1981, *Inter-rater and intra-rater reliability of the oral interview and concurrent validity with cloze procedure*, in Palmer, A.S., Groot, P.J.M., and Trosper. G.A., eds.

Shohamy. E., 1983, *Inter-rater and intra-rater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew*, in Oller, J. W., ed.

Stansfield. C.W., and Kenyon, D.M., *Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview*, System, 20.

Underhill, N. 1987. *Testing Spoken Language: A Handbook of Oral Testing Techniques.* Cambridge: Cambridge University Press.

van Lier, L., 1989, *Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation*, TESOL Quarterly, 23.

Wolfson. N., and Judd, E., eds., 1983, *Sociolinguistics and language acquisition*, Newbury House, Rowley, Massachusetts.

Wright, B.D. and Masters, G.N., 1982, *Rating scale analysis: Rasch measurement*, Mesa Press, Chicago.

Young. R.F. and Milanovic, M., 1992, *Discourse variation in oral proficiency interviews*, Studies in Second Language Acquisition, 14.

Appendix 2: **Criteria for assessment**

Fluency

This rates the naturalness of the speed and rhythm together with lack of hesitation and pauses. Pauses for thought rather than language should be regarded as natural features of spoken interaction and not penalised.

Accuracy and range

Range is the quantity and correctness, the quality of both grammatical structures and vocabulary. The major errors are those which obscure the message, they should be penalised more heavily than minor ones, those that do not obscure the message. Obvious slips of the tongue should not be penalised.

Pronunciation

This covers both individual sounds and the pronunciation in a stream of words, that is stress, timing, rhythm, pauses, intonation patterns and range of pitch within utterances. It is not expected that candidates' pronunciation should be entirely free of L 1 features.

Task achievement

This scale measures a candidate's participation in the four phases of the speaking paper and covers the following areas: .

- appropriacy and relevance of contributions to the tasks; .
- independence in carrying out the tasks set, that is the degree to which each candidate can carry out the tasks without prompting or redirection by the interlocutor or the other candidate;
- the organisation of the candidate’s contributions (logical or coherent sequencing of utterances)
- the candidates' flexibility and resourcefulness;
- the degree to which the candidates' language contributes to successful task management, to the selection of appropriate language functions and vocabulary.

Interactive communication

This refers to the candidate's ability to interact actively and responsibly, with sensitivity to the norms of turn-taking appropriate to each phase of the test. Candidates who are unwilling or unable to take their turn adequately will receive a reduced score on this scale.

These are the scales, usually placed in blocks of two:

- 0 : candidates not present or not producing enough language to assess;
- 1 and 2 are clear failures: a candidate well below the standard that is expected in the exam;
- 3 and 4 are on the down side of the border line, candidates who are approaching the right level but are below the level expected;
- 5 and 6 are on the plus side of that borderline, candidates of pass level;
- 7 and 8 are candidates of well above pass level.

The scales descriptors are fairly clear, as shown in the following chart.

Appendix 3: **Scales used in the Cambridge CAE examination**

Cambridge Certificate in Advanced English - Paper 5
Speaking Criteria for Assessment

FLUENCY	ACCURACY AND RANGE	PRONUNCIATION	TASK ACHIEVEMENT	INTERACTIVE COMMUNICATION
7 - 8 Coherent spoken interaction with speed appropriate to the task and few intrusive hesitations	**7 - 8** Evidence of a wide range of structures and vocabulary in all contexts. Errors minimal in number and gravity	**7 - 8** Little L1 accent/L1 accent not obtru-sive. Competent handling of most English pronuncia-tion features.	**7 - 8** Tasks are dealt with fully and effectively, with notable coherence and organisation of salient points. The language is fully appropriate to each task.	**7 - 8** Contributes fully and effectively throughout the interaction, with sensitiveness to the norms and require-ments of turn-taking in each task.
5 - 6 Occasional but noticeable hesita-tions, but not such as to strain the listener or impede communication. Pauses to marshal thoughts rather than language.	**5 - 6** Evidence of appropriate range of structures and vocabulary; has the range needed to express inten-tion. Number and gravity of errors do not impede the message.	**5 - 6** Noticeable L1 accent with minor difficulties with several features. These cause only isolated strain or incomprehension and do not impede communication or comprehension.	**5 - 6** The tasks are dealt with effectively, but treatment may be fragmented or a little unsystematic. The language is generally appro-priate, with only isolated lapses.	**5 - 6** Contributes with ease for most of the interaction with only occa-sional and minor difficulties in negotiation or turn-taking.
3 - 4 Fairly frequent and noticeable hesitations. Communication is achieved but strains the listener at times. May need to pause to marshal language.	**3 - 4** Fairly frequent errors and evi-dence that range of structures and vocabulary limits full expression of intent. Communi-cation of the essential message is not prevented.	**3 - 4** Obvious L1 accent with major defects in some areas. These may frequently strain the listener and/or make compre-hension of detail difficult.	**3 - 4** One or more of the tasks dealt with in a limited manner. The language is noticeably inappropriate at several points. Redirection may have been required at times.	**3 - 4** Contributes effectively for much of the interaction, but with intrusive difficulties or deviations at times. Responses may be short, without attempt at elaboration.
1 -2 Disconnected speech and/or frequent hesitations impede communication and constantly strain the listener.	**1 - 2** Frequent basic errors and limited range of structures and/or vocabulary impede communi-cation of the essential message and constantly strain the listener.	**1 - 2** Heavy L1 pronun-ciation and wide-spread difficulties with English features impede communication of the basic message and constantly strain the listener.	**1 - 2** Inadequate or irrelevant attempts at the tasks with much inappro-priate language. Requires major or repeated redirec-tion or assistance with the tasks.	**1 - 2** Difficulty in maintaining contributions throughout. May respond to simple or structured interaction but obvious limitations in freer contexts.
0	Inadequate for assessment, even after prompting by the interlocutor.			

UCLES, October 1991