Les Après-midi de LAIRDIL

The Problems of Oral Testing. What Did you Say?

par

Mike NICHOLLS

et

Contributions de LAIRDIL

Conférence n° 1
25 novembre 1993

LAIRDIL

IUT A
115 route de Narbonne
31 077 Toulouse Cédex

Tél.: 62 25 80 43

_La_boratoire _I_nter-Universitaire de _R_echerche en _Di_dactique des _L_angues

Aimée Blois, Bernard Crosnier, Nicole Décuré,
Françoise Lavinal, Anne Péchou, Christine Vaillant

Créé en 1989, LAIRDIL est un laboratoire inter-universitaire de recherche de l'Université Toulouse III et de l'INSA, rattaché à l'IUT A. Il a pour objet la recherche en didactique des langues. Les travaux sont principalement axés sur l'étude de l'apprentissage des langues en autonomie. La diffusion des résultats de cette recherche est une priorité.

Chaque année, LAIRDIL organise donc un cycle de séminaires-conférences sur des sujets de pédagogie ou de didactique susceptibles d'intéresser un grand nombre d'enseignant(e)s d'anglais, voire d'autres langues. La conférence enregistrée est transcrite. Les membres du laboratoire ajoutent leurs réflexions propres sur le thème abordé. Le tout est édité dans une brochure.

La conférence n° 1, *The Problems of Oral Testing. What Did you Say?***,** a été donnée le 25 novembre 1993 par Mike Nicholls, responsable régional Midi-Pyrénées des examens d'anglais comme langue étrangère de l'Université de Cambridge depuis 1989.

Dans sa conférence, Mike Nicholls fait référence aux travaux de recherche et au modèle proposé par Mike Milanovic et Nick Saville de University of Cambridge Local Examinations Syndicate (UCLES).

Autres séminaires :

- *Autonomous Learning of Vocabulary Through Extensive Reading*, par James Coady, 27 janvier 1994 .
- *Rereading Video*, par Richard Cooper, 31 mai 1994.
- *Fluency and Appropriacy in Oral English* par J.D. Brown, 16 janvier 1995.
- *Maximizing the Value of Jigsaw Activities*, par Stephen Gaies, 6 mai 1995.

Décryptage de la conférence : Aimée Blois et Bernard Crosnier
Réalisation de la brochure : Nicole Décuré

SOMMAIRE

CONFERENCE

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.

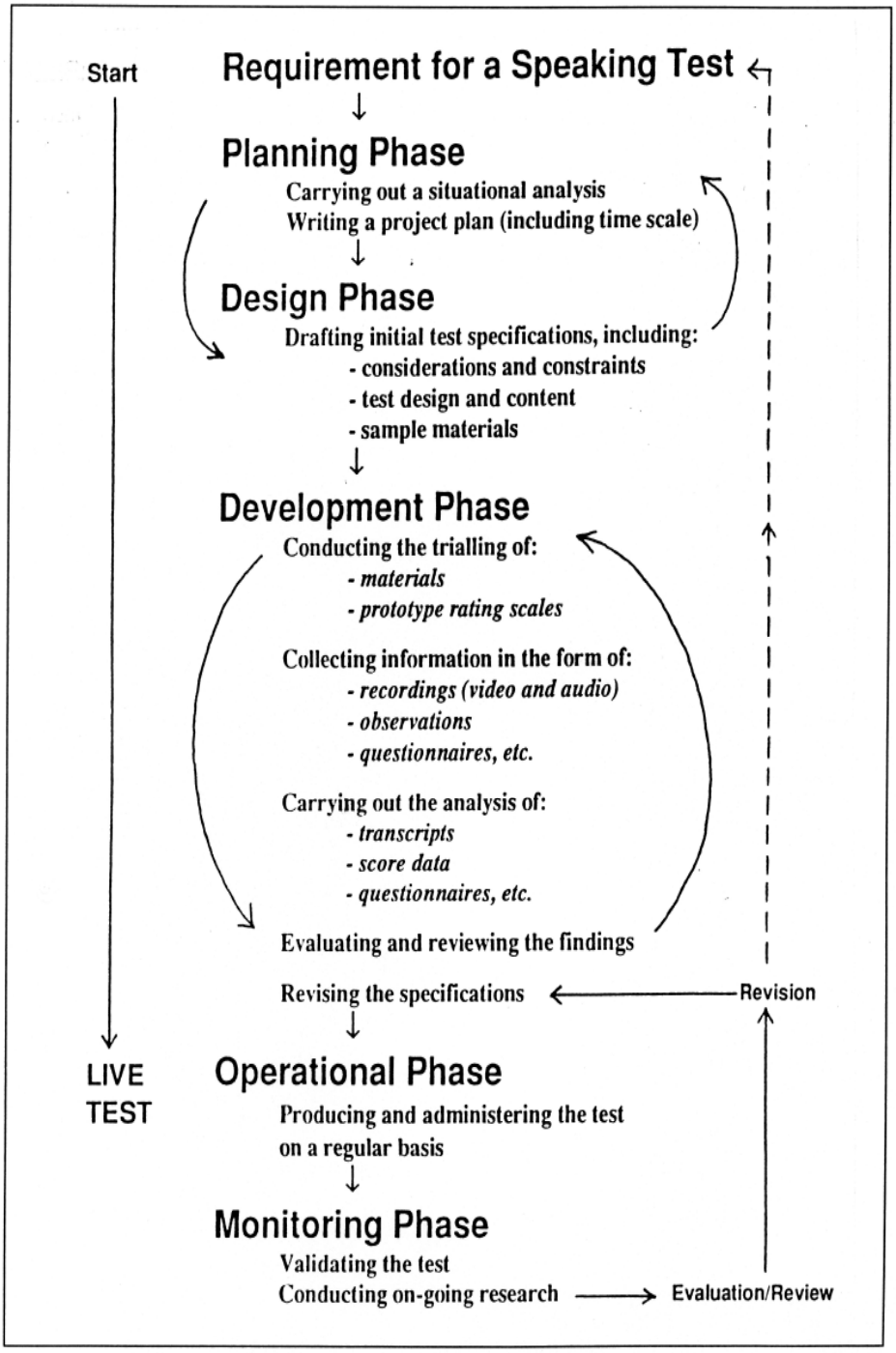
After the sample materials have been prepared, the *development phase* begins in which the trailing of the sample materials and of the prototype rating scales is undertaken. Information is collected on the efficacy of the test and its procedures, usually in the form of audio and video recordings of what 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

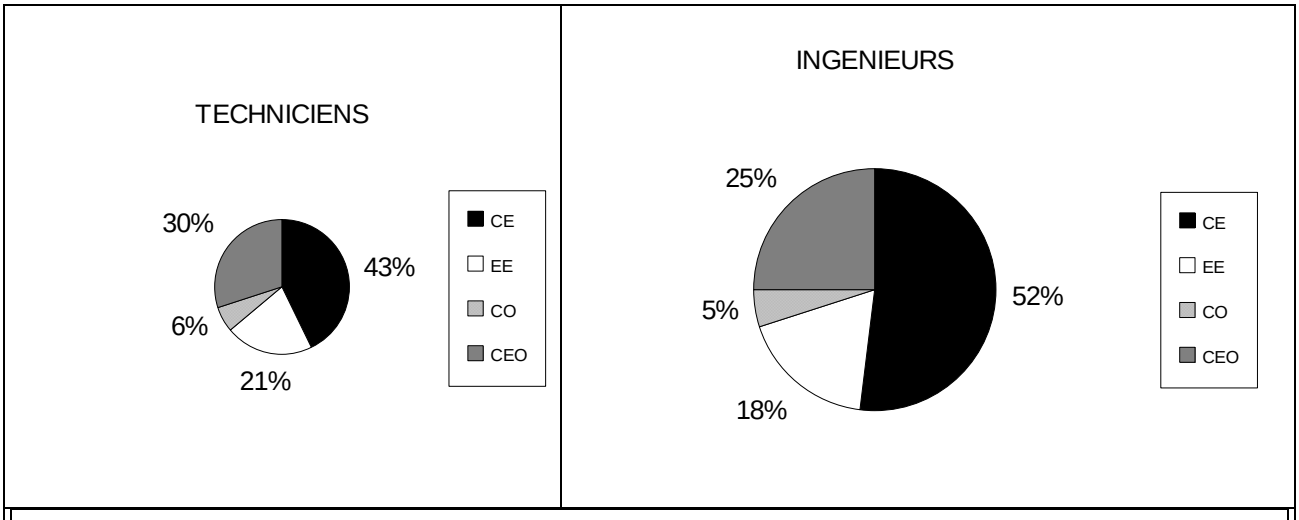
CONTRIBUTIONS DE LAIRDIL

Besoins langagiers et pertinence des tests

Bien que l'importance de la production orale par rapport aux autres aspects de la langue ne soit pas le propos de la conférence faite par Mike Nicholls, il n'est peut-être pas inutile de faire quelques remarques au sujet de ce qu'il a dit:

> The testing that we do most of the time is written testing, which is virtually irrelevant of what people normally do. 90% of our students, when they leave their institutions, will not use the language in its written form or very rarely. They will do almost no writing in English after they leave and they will not do a great deal of reading probably but they will have to listen a lot, if they are using the language at all. (...) Oral language is more important in everyday use than the written language, as it is with the native language.

On pourrait, me semble-t-il, reprocher à ce commentaire de n'être pas assez nuancé. Si la langue orale est en effet beaucoup plus importante que la langue écrite dans la vie courante, ce n'est cependant pas toujours le cas dans une activité professionnelle. On peut donner pour exemple et pour preuve les résultats d'une enquête auprès des techniciens supérieurs titulaires d'un DUT ou occupant une fonction à laquelle peuvent prétendre nos diplômés et, pour comparaison, à un certain nombre d'ingénieurs des sociétés suivantes: Aérospatiale (Toulouse), Airbus Industrie (Toulouse), Alsthom Atlantique (Belfort), Bosch (Rodez), EDF (Paris), Ford (Bordeaux), Forestline (Capdenac), GES Barrouy (Saint Jory), Hewlett Packard (Grenoble), Ravina (Toulouse), Renix (Toulouse), Single Buoy Moorings (Monaco), SMC (Castres), SNIAS (Méaulte), SOCATA (Tarbes), SOGERMA (Bordeaux), Soule (Bagnères de Bigorre), TELECOM (Paris), Thomson-CSF (Toulouse).



Répartition des différentes activités langagières

Les diagrammes ci-dessus montrent, pour les techniciens supérieurs (TS) d'une part et pour les ingénieurs (ING) de l'autre, comment se répartissent la compréhension de la langue écrite (CE), l'expression écrite (EE), la compréhension de la langue parlée (CO) et l'expression orale qui implique généralement la compréhension (CEO).

Contrairement à ce que pourrait laisser penser la généralisation faite par Mike Nicholls, c'est la langue écrite qui domine, tant chez les techniciens supérieurs où elle représente 70% des activités en langue étrangère, que chez les ingénieurs. La production orale n'est toutefois pas négligeable et à l'intérieur de cette compétence la répartition plus fine des types d'activité donne les résultats suivants:

- 17% de compréhension seulement à l'occasion d'une réunion ou d'une conférence à laquelle les intéressés assistent sans intervenir;
- 6% d'expression en tant que conférencier;
- 29% de conversation pour l'accueil de collègues ou de clients étrangers;
- 15% de négociation;
- 4,5% de formation de clients étrangers;
- 28,5% de conversation téléphonique.

Ces chiffres sont très intéressants en rapport avec ce que Mike Nicholls dit de la validité des tests oraux. 77% des activités mentionnées ci-dessus impliquent un face-à-face avec, le plus souvent, son homologue étranger. Si l'on veut faire passer des tests qui se rapprochent le plus possible de ce que les gens font dans la réalité, c'est manifestement le modèle préconisé par Nicholls, à savoir le dialogue entre deux candidats, qui est le plus satisfaisant. Faire parler un candidat sur tel ou tel sujet, lui faire décrire une image, etc., s'apparente plus à la situation d'un conférencier, qui ne représente que 6% des activités. Encore peut-on noter qu'il est plutôt rare que ce dernier n'ait aucun échange avec des auditeurs, par exemple après son intervention. Il est d'ailleurs à remarquer qu'il en va de même dans la vie courante. Quand se trouve-t-on réellement dans une situation où l'on ne fait qu'écouter ou que parler? Tester la compréhension orale en dehors de l'expression et l'expression sous forme de monologue pourrait donc encourir le reproche d'être trop artificiel, même si cela peut constituer de bons exercices d'entraînement. Le test en binômes a l'avantage de ressembler plus à une situation réelle, d'être plus naturel donc, sans doute, moins stressant, comme cela a été souligné, et surtout peut-être, plus pertinent.

Aimée Blois

L'épreuve orale de langues vivantes aux concours communs polytechniques

Recrutement dans les écoles nationales supérieures d'ingenieurs - ENSI

Présentation de l'épreuve orale

Le candidat au concours d'entrée des Ecoles Nationales Supérieures d'Ingénieurs (ENSI) entend trois fois, à quelques secondes d'intervalle, un enregistrement de quatre minutes environ en laboratoire. Il prend des notes pendant l'écoute qui n'est jamais interrompue. Il dispose ensuite de dix à douze minutes pour organiser et mettre en forme son résumé, assorti d'un commentaire sur un ou plusieurs aspects du sujet qu'il sélectionne lui-même. Le candidat doit organiser ses notes le mieux possible car le temps dont il dispose ne lui permet pas de rédiger sa présentation.

- Les trois écoutes plus le temps de préparation à l'épreuve durent de 25 à 30 minutes.

- La présentation orale individuelle et la conversation durent environ 25 minutes en face à face avec l'examinateur.
- L'examinateur dispose de cinq minutes pour rédiger son rapport et noter le candidat suivant une grille d'évaluation (*Document 1*).

Sujets proposés a l'oral du concours

a) *Sujets*

Les sujets sont enregistrés sur des cassettes audio à partir d'extraits d'articles de la presse écrite dont les références (titre, date, publication) sont indiquées sur l'enregistrement. Les sujets traitent de tous les problèmes de société, à l'exclusion de ceux qui pourraient prêter à violente polémique (politique, religion, violence, etc.). Ils sont fournis par chacun des examinateurs retenus pour l'oral du groupe "Concours Communs Polytechniques". Ils doivent être amendés avant l'enregistrement, si nécessaire, dans le but d'éviter les difficultés et complexités excessives dans le domaine lexical, étant donné le temps limité dont dispose le candidat pour écouter et préparer son épreuve. Les textes publiés dans les ouvrages et magazines couramment utilisés dans les classes préparatoires aux grandes écoles et universités sont exclus de la

sélection du jury, même s'ils correspondent bien aux besoins définis pour l'oral du concours écoles d'ingénieurs.

b) *Les enregistrements sur cassette*

La réalisation des cassettes est effectuée par des anglophones à partir d'un enregistrement à la vitesse "normale" d'une lecture courante. L'accent du lecteur ou locuteur est en principe britannique, celui-ci étant le plus familier à la majorité des candidats du concours.

Limites et imperfections de l'oral du concours

Le sujet idéal serait un texte à plusieurs voix, enregistré en direct : conversation ou débat authentique. Mais la nécessité de préparer un grand nombre de sujets de qualité technique et de difficulté à peu près analogues limite considérablement les ambitions du jury en ce domaine. Chaque année, tous les documents sont renouvelés, ce qui constitue une difficulté supplémentaire à prendre en considération. Pour la session d'oral 1994, le jury a retenu 56 sujets sur plus de 150 proposés. De plus, les examinateurs savent bien que la préparation de l'oral en classes préparatoires aux grandes écoles se fait essentiellement à partir de textes de la presse écrite.

S'il apparaît clairement que l'écrit ainsi oralisé n'est pas une solution idéale, la formule choisie correspond à un dénominateur commun à tous les candidats et permet donc d'harmoniser au maximum l'oral de l'épreuve de langues vivantes tout en garantissant une certaine densité du contenu.

Evaluation des performances

Tous les examinateurs utilisent une fiche type évaluant les compétences et la prestation de chaque candidat.

Pour 50% de la note : ***compréhension et production***

- Compréhension du sujet, idées principales, cohérence du résumé et synthèse du document.
- Pertinence et richesse du commentaire, aptitude à présenter clairement une analyse.

Pour l'autre moitié de la note : ***qualités de la langue parlée***

- Compétence dans le domaine de la correction grammaticale, richesse syntaxique, lexicale, morpho-syntaxique.

- Performance du point de vue du registre phonologique, c'est-à-dire accentuation, intonation et rythme, correction du schéma vocalique et consonantique, sons individuels et dans une suite de mots.

Pour la session 1993, la moyenne de l'épreuve orale du concours était 10,81 avec un écart-type de 3,57.

Pour la session 1994, 3002 candidats ont été interrogés à l'oral sur un total de 6785 aux épreuves écrites. La moyenne est de 10,89 avec un écart-type de 3,445 (*Document 2*).

Entraînement des candidats : informations aux enseignants des classes préparatoires et examinateurs

En raison du temps relativement bref laissé au candidat pour préparer son épreuve orale de langue vivante, soit environ quinze minutes après l'écoute au laboratoire, seuls un entraînement régulier et un encadrement rigoureux permettent d'offrir le niveau de prestation et de compétence attendu par le jury. La compréhension ponctuelle du texte ne garantit pas à elle seule l'obtention d'une note satisfaisante.

L'oral du concours demeure "un exercice de communication et d'expression valorisant également l'aptitude des futurs ingénieurs à réagir intellectuellement de façon autonome, à structurer leurs pensées et à se montrer curieux autant qu'informés des événements et des problèmes de société contribuant à l'évolution du monde actuel"
(Extraits des rapports et notes fournis aux examinateurs du concours et aux professeurs de langues des classes préparatoires aux grandes écoles).

Bernard Crosnier

Document 1

CONCOURS COMMUNS POLYTECHNIQUES LANGUES VIVANTES
GRILLE D'EVALUATION

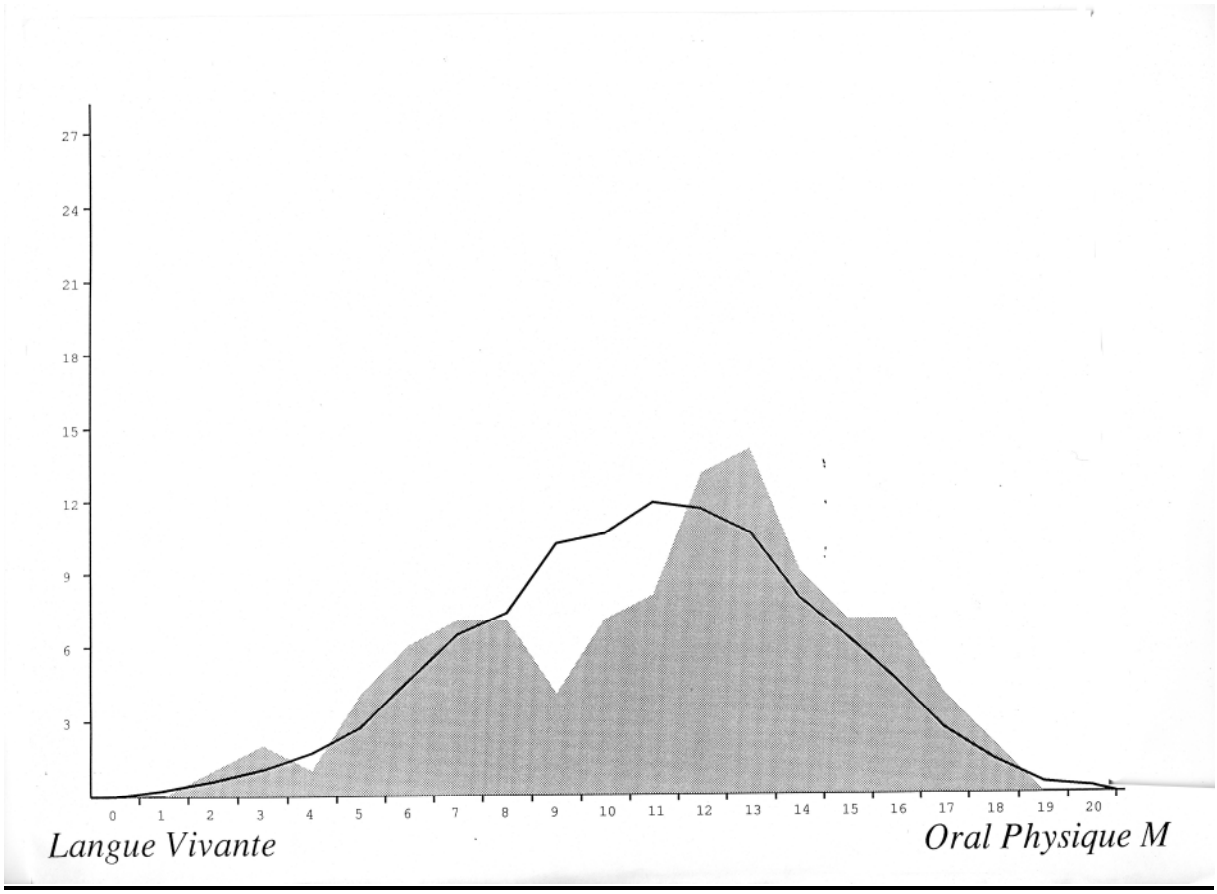
Date : **Nom :** **NOTE /20** ❍
Heure : **Prénom :**
Sujet :

	Note /20	Coeff	Total
1. Compréhension du texte • Idées principales • Cohérence du résumé • Compréhension des propos de l'examinateur		2	
2. Commentaire du texte • Contenu • Pertinence (information, culture, réflexion) • Cohérence (organisation) • Aptitude à dialoguer		3	
3. Contenu grammatical et syntaxique • Maîtrise des points-clés (détermination, aspect, temps, etc.) • Variété des structures • Aptitude à se corriger spontanément		2	
4. Contenu lexical • Vocabulaire de base connu? • Vocabulaire spécifique au sujet connu? • Gallicismes/barbarismes (pénalisation) ou idiotismes (bonification)		2	
5. Contenu phonologique • Intelligibilité globale (y compris rythme et débit) • Accentuation (mot, phrase) • Sons déformés compromettant l'intelligibilité • Intonation (plus ou moins francisée) • Interruptions en français		1	

Appréciation générale / remarques : Total /200 ☐

Note /20 ❍

Document 2 : **Répartition des notes de l'examinateur *x***
Tableau comparatif avec la moyenne des correcteurs



Un test, ça se teste
Analyse d’un test de niveau

Buts de l’étude

Au cours de la discussion qui a suivi la conférence, Mike Nicholls a affirmé que les résultats des tests écrits de Cambridge n’étaient pas différents, de façon significative, de ceux d’un test oral. “ The results of the oral correlate remarkably well with the results of the written test. ” Pourquoi alors faire un test écrit, argumente-t-il, puisque nos cours sont, dans la plupart des cas, essentiellement axés sur l’oral? Mais pourquoi faire un test oral, pourrait-on rétorquer? Car tout test d’expression orale prend beaucoup de temps et il est donc difficilement compatible avec de grands nombres.

Il faut d’abord savoir à quoi sert un test. Sanctionne-t-il un travail ou est-il un test d’évaluation d’un niveau de compétences avant de commencer un cours ou une formation? Le contexte est bien différent. Dans le premier cas, l’apprenant(e) peut se sentir frustré(e), trompé(e) d’avoir travaillé l’oral et de n’être évalué(e) que sur l’écrit. Mais dans le deuxième cas, procéder à un test oral devient une question de temps, donc de moyens.

Cette remarque, non développée, de Mike Nicholls sur la corrélation de l’écrit et de l’oral m’a amenée à examiner le test décrit ci-dessous, utilisé par mes collègues et moi-même depuis une vingtaine d’années pour classer les étudiants de second cycle scientifique et médical de l’Université Paul-Sabatier à Toulouse en groupes de niveau en début d'année. Il est bien loin de remplir toutes les conditions énumérées dans la conférence et cependant il nous donne satisfaction car les classements ainsi opérés s'avèrent fiables à quelques exceptions près. Mais, alors qu'il a été construit de façon empirique, il paraissait intéressant d'examiner son fonctionnement, ses avantages et ses limitations. Cette étude a permis de soulever quelques questions qui méritent d'être examinées dans un travail de recherche ultérieur.

Contexte

Le test s'adresse à des étudiants qui vont suivre 75 heures de cours à raison de trois heures par semaine, subdivisées en une séance axée sur l'écrit (grammaire, compréhension et expression écrites) et une partie axée sur l'oral

(compréhension et expression orales). En fait, les deux parties accordent une place prépondérante à la communication orale.

Bien que le test ait été raccourci au fil des ans, sa conception est restée la même ainsi que son contenu, ce qui présente l'avantage de donner des résultats consistants d'une année à l'autre. Il teste la compréhension écrite, les connaissances grammaticales, syntaxiques, lexicales et la compréhension orale. L'expression orale a été volontairement écartée car elle prend trop de temps et ce test doit être administré à beaucoup d'étudiants à la fois (700 à l'heure actuelle) en un temps très court. Beaucoup d'exercices sont sous forme de QCM qui permettent une correction rapide et quelques-uns exigent des réponses plus ouvertes. Le test se fait en temps limité (les plus faibles ne finissent pas) et la partie de compréhension orale se déroule de façon identique à toutes les séances.

Ce test est insatisfaisant sur le plan intellectuel car il ne teste qu'une petite partie des connaissances et compétences des étudiant(e)s, notamment en n'évaluant pas la communication orale. Mais il est satisfaisant sur le plan pratique car le barème établi nous permet de classer les étudiant(e)s en quatre niveaux (de débutant à avancé) avec très peu d'erreurs d'appréciation. Les erreurs de classement sont dues essentiellement à des erreurs d'addition (!), au copiage, à la mauvaise forme le jour du test ou, au contraire, à un hasard heureux. Il y a peu de changements de niveau par la suite, 1 à 2%, essentiellement les cas limites. Ces changements se font surtout vers le bas, à l'exception du niveau 1, les faux-débutants : ils font souvent un mauvais test parce qu'ils/elles n'ont pas fait d'anglais depuis longtemps et il suffit de quelques semaines pour réactiver leurs connaissances. Depuis que les niveaux sont équivalents pour l'obtention d'une u.v. libre (ce qui n'était pas le cas au début où seul le niveau 2 donnait l'u.v.) il y a moins de tricherie car il n'est pas dans l'intérêt des étudiant(e)s d'être admis(es)dans un niveau trop fort. On peut aussi constater avec plaisir que très peu font délibérément un test plus faible afin de se trouver dans un groupe qui sera facile. C'est une u.v. libre, donc choisie, et la grande majorité a envie de progresser et non simplement d'obtenir un diplôme. De plus, le niveau atteint étant mentionné sur le diplôme, il vaut mieux, pour son CV, avoir un bon niveau. Enfin, bien que le test privilégie l'écrit, dans la grande majorité des cas, les étudiant(e)s sont à l'aise dans la partie orale du cours, mais en général, plus faibles, surtout au niveau 4.

Matériau

Sur les 700 tests d'octobre 1993, ont été gardés, pour faciliter les calculs :

- *niveau 2* : 200 tests (17 éliminés : tests écrits seulement, changements de niveau, non-inscrits, erreurs d'addition;
- *niveau 3* : 200 tests (les premiers dans l'ordre alphabétique);
- *niveau 4* : 150 tests (moins d'étudiant(e)s, les résultats ont été reportés sur 200);
- Les résultats du niveau 1 n'ont pas été pris en compte car le nombre de copies était trop faible.

Méthodologie

Toutes les notes ont été entrées dans une base de données : pour chaque étudiant(e) les notes de chaque test, les notes totales de l'écrit, de l'oral et l'addition.

Il a été effectué

- des graphiques comparatifs à l'intérieur d'un même niveau : entre divers exer-cices, entre écrit et oral en tenant compte des totaux d'exercices (ramenés à 10);
- des graphiques comparatifs entre niveaux (sur même base de nombre);
- un calcul de l'écart entre écrit et oral.

Description des exercices

Compréhension et expresion écrites (total des points : 72)

	Note	Objet	Type	Tâche
Ex. 1	10	CE : reconnaître les verbes dans des titres de journaux	lecture, fonction des mots	recopier un mot
Ex. 2	10	Vocabulaire. : mots outils, expressions de temps (synonymes)	QCM	cocher
Ex. 3	6	questions : formulation	écrire des phrases	écrire
Ex. 4	9	adverbes : liste , mettre des adverbes dans des phrases	exercice lacunaire	choix dans liste
Ex. 5	5	structure de la phrase, formes verbales	éléments dans le désordre	flèches
Ex. 6	10	emploi des temps dans des phrases séparées	QCM	cocher
Ex. 7	12	accord des temps dans un texte suivi	transformation de l'infinitif	écrire
Ex. 8	10	trouver des erreurs variées dans des phrases et corriger	lecture, correction	écrire

Compréhension orale (total des points : 42)

	Note	Objet	Type	Tâche
Ex. 9	10	discrimination des sons	reconnaissance de sons	écrire des chiffres
Ex. 10	10	compréhension de mots et de leurs définitions	QCM oral	cocher
Ex .11	10	reconnaître un mot dans une phrase	QCM oral	cocher
Ex. 12	12	compréhension de phrases (stéréotypes culturels) à l'aide de dessins	compréhension globale	écrire des chiffres

Deux questions principales ont été examinées :

- *Peut-on raccourcir le test ?*

L'augmentation des effectifs rend la correction de plus en plus fastidieuse. Le test, dans sa forme actuelle, n'est pas informatisable. Certaines questions, aussi fermées soient-elles, admettent un grand nombre de réponses, d'où la nécessité d'une correction "intelligente". L'informatisation d'ailleurs, en ne laissant que le seule possibilité de QCM, rend le test moins riche car il n'évalue que les connaissances passives et non actives, la réception et non la production.

Analyser les résultats du test permettrait de voir si certains exercices sont redondants, donnant les mêmes résultats à l'intérieur d'un même niveau. Par exemple, nous avions l'impression subjective que les deux premiers exercices, bien que portant sur des compétences très différentes, donnaient les mêmes notes sur les copies.

Deux ou trois exercices peuvent-ils être suffisamment révélateurs, malgré la notion, couramment admise, que plus un test est long, plus il est fiable (Hughes, 36) ?

Y a-t-il des exercices qui donnent des résultats plus significativement différents d'un niveau à l'autre ? Il nous semblait que c'était le cas de certains exercices, en particulier ceux portant sur la formulation de questions et l'accord des temps.

- *Y a-t-il concordance entre la partie écrite et la partie orale ?*

La partie écrite peut se faire individuellement mais la partie orale nécessite un magnétophone, donc une personne présente qui veille à ce que le test se déroule de la même façon pour tout le monde. Lorsque des étudiant(e)s

doivent passer le test en retard, ils/elles ne font que la partie écrite. Ces résultats aussi donnent satisfaction. La compréhension orale est-elle alors redondante ? La supprimer permettrait de gagner du temps et de se débarrasser d'exercices dont la correction est difficile.

- Il était également tentant d'essayer de voir si les niveaux de départ se retrouvent à l'arrivée. En fait, trop de paramètres entrent en ligne de compte qui sont difficilement mesurables : le travail fourni par l'étudiant(e), les professeurs différents, les notations différentes.

L'étude est loin d'être terminée mais nous présentons ici les questions qui se sont posées en cours d'étude et quelques conclusions.

Remarques sur les exercices

Comme il est dit plus haut, les exercices ne sont pas que des QCM et laissent donc place au choix individuel, à la production (même limitée) de langage, aux compétences actives plutôt que passives. Il n'y a pas de hasard possible, comme dans un QCM. L'exercice donnerait peut-être des résultats très différents si des choix étaient proposés, les étudiant(e)s sachant plus facilement reconnaître une phrase juste que la produire. Cette étude sera menée ultérieurement.

Tout test présente des difficultés qui n'ont rien à voir avec les compétences langagières :

- *Les difficultés inhérentes aux types d'exercices proposés* : l'exercice 3, très artificiel, nécessite des exemples en français pour que l'on comprenne bien ce qui est demandé ; la tâche des exercices oraux nécessite aussi des explications et démonstrations.
- *Une évaluation du temps nécessaire et une juste répartition* : si les exercices de la page 1 donnent de meilleurs résultats que ceux de la page 3, ce n'est pas qu'ils sont plus faciles (l'exercice 1 est assez difficile) mais les étudiant(e)s y consacrent sans doute plus de temps, car c'est le début de la séance et ils/elles évaluent mal le temps global nécessaire. S'ils étaient en fin de test, les résultats seraient peut-être modifiés. C'est également une étude à mener.
- *Un temps d'acclimatation au test*, notamment pour le test oral. Certain(e)s n'ont pas entendu d'anglais depuis longtemps.

- Pour les tests oraux, des éléments tels que la mémoire interfèrent, notamment dans les tests 10 et 11 où quasiment personne ne pense à noter le premier mot, ce qui aiderait grandement dans l'accomplissement de la tâche.
- Pour le test 9, une bonne oreille permet de pallier les déficiences de connaissances. Différencier un mot d'un autre n'implique pas qu'on sache ce qu'ils veulent dire et encore moins les employer. C'est aussi un test qui requiert d'avoir bien compris la technique (deux chiffres sous un mot entraînent les deux mêmes chiffres sous l'homophone) et exige des réflexes rapides.
- Enfin, les références culturelles peuvent faire défaut à certain(e)s : c'est le cas du test 12 qui porte sur des stéréotypes de nationalité. On peut aussi ajouter un paramètre pour cet exercice. Le dessin d'origine était moins clair. Depuis que le dessin a été refait, les résultats sont meilleurs. Ce n'est pas la compréhension qui s'est améliorée, mais la lisibilité du dessin.

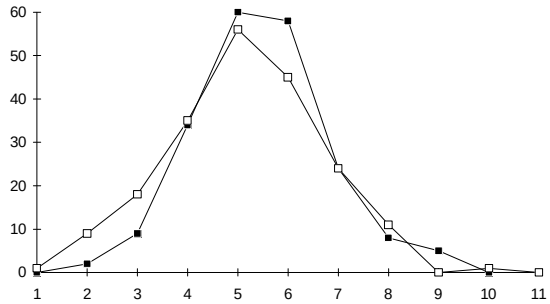
Quelques résultats

Essayons de répondre aux questions de départ.

- *Peut-on raccourcir le test?*

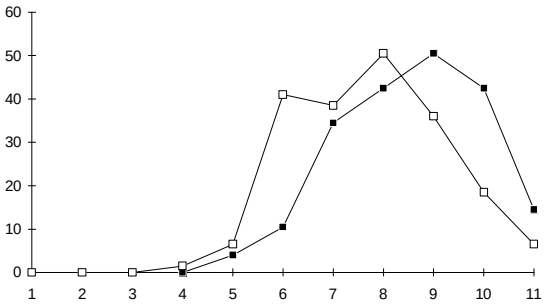
L'impression de départ, selon laquelle les exercices 1 et 2 donnent des résultats quasiment identiques est confirmée.

Niveau 2



Exercices 1 et 2

Niveau 4



Exercices 1 et 2

COMPARAISON EXERCICES 1 ET 2

A la lecture des moyennes de notes de chaque exercice, on s'aperçoit que le test 9, déjà mentionné, n'est pas " rentable ", car c'est le test pour lequel il y a le moins de différence d'une copie à l'autre. Comme c'est aussi le plus long et le plus difficile à corriger, nous avons pu le supprimer cette année sans regret.

	n. 2	**n. 3**	**n. 4**
exercice 1	4.5	5.7	7.6
exercice 2	4.2	5.3	6.7
exercice 3	1.9	3.7	4.3
exercice 4	2.8	4.3	5.7
exercice 5	2.6	3.5	4.3
exercice 6	3.7	4.8	6.7
exercice 7	3.1	5.1	7.4
exercice 8	0.4	1.2	2.9
total écrit	23	33	45.7
total sur 20	***6,4***	***9,2***	***12,7***
exercice 9	*3.6*	*4.3*	*5.7*
exercice 10	2.8	3.7	4.9
exercice 11	4.5	5.2	6.8
exercice 12	5.9	7.1	8.6
total oral	18.3	20.3	26
total sur 20	***8,7***	***9,6***	***12,4***
total général	39.8	57.8	71.7

MOYENNE DES NOTES

- *Y a-t-il concordance entre la partie écrite et la partie orale ?*

Les notes ramenées sur 20 viennent confirmer ce que Mike Nicholls affirmait: la corrélation entre les résultats d'écrit et ceux d'oral, sauf pour les niveaux faibles. La supériorité de l'oral, dans ce cas, peut s'expliquer par la nature même des tests qui, comme dit plus haut, requièrent une bonne oreille et une bonne mémoire plus qu'une réelle compréhension.

Lorsque le test est administré individuellement, seule la partie écrite est donnée et un barème a été établi. En fait, pour affiner l'analyse, si l'on regarde uniquement les notes de la partie écrite, de façon individuelle et non plus globale, on s'aperçoit que la note d'oral est loin d'être négligeable car elle corrige la note d'écrit, surtout dans le niveau 3.

- En effet, au niveau 4, le nombre d'étudiant(e)s ayant obtenu moins de 37 points, la barre fixée par extrapolation, est infime: 10 sur 150, dont 5 ont une note globale à la limite des niveaux 3 et 4.
- Au niveau 2, 43 étudiant(e)s sur 200, soit un peu moins d'un quart, ont une note d'écrit très faible. Parmi eux/elles, une dizaine est à la limite du niveau 1 et

six ont une note d'oral nettement supérieure. Pour le reste, la compréhension orale a compensé.

- Pour un tiers du niveau 3, la partie écrite est soit au-dessous, soit au-dessus des barres fixées.

Doit-on en conclure que pour économiser du temps on pourrait, comme dans les oraux de rattrapage, ne pas faire de test oral pour ceux et celles qui ont une note élevée à l'écrit? Le test oral permettrait d'affiner la notation pour les niveaux les plus bas.

Conclusion

Le test décrit ci-dessus est loin de remplir les critères de validité et de fiabilité qui semblent nécessaires à une évaluation objective et fine. Mais, *in fine*, tout test ne vaut que par l'habitude qu'ont les enseignant(e)s concerné(e)s de l'utiliser car ils/elles l'ajustent à leurs besoins, leur connaissance du terrain, leur enseignement.

Nicole Décuré

Une gageure : tester l'oral

La manière d'évaluer l'anglais en tant que langue étrangère a beaucoup évolué au cours des dernières années, comme en témoigne l'évolution des pratiques à l'INSA de Toulouse. L'évaluation de la performance orale, en particulier, a pris une place accrue dans la notation globale. Inspiré par l'une des épreuves de l'examen de Cambridge First Certificate, l'examen oral en première et deuxième année - 400 étudiants environ - prend la forme d'un jeu de rôles ou d'une discussion où deux étudiants doivent résoudre un problème. L'examinateur reste en retrait et n'intervient pas.

Plusieurs années de ce type d'évaluation ont inspiré les réflexions suivantes.

Choix du type de test

- *Un test doit refléter ce qu'on enseigne*. Hughes use le terme de *backwash* pour designer l'effet d'un type de test sur l'apprentissage et sur la façon d'enseigner. Il serait vain d'insister sur la pratique de la langue de communication en cours si on persistait à évaluer par une épreuve écrite. L'apprenant est peu enclin à fournir in effort à l'oral s'il sait qu'il ne sera pas évalué à l'oral.
- *Il doit être discriminant*, c'est-à-dire permettre d'utiliser une vaste gamme de notes.
- *Il doit tenir compte des conditions matérielles*: temps, locaux et personnel disponible.

Durée de l'épreuve

Quinze minutes en moyenne pour deux étudiants avec un seul examinateur, faute de moyens pour en avoir deux, soit huit étudiants à l'heure. Le choix des "paires" est laissé aux étudiants. Chaque groupe est convoqué un quart d'heure à l'avance et se prépare dans la salle où l'épreuve est soutenue.

Problèmes rencontrés

La difficulté principale est d'assurer l'harmonisation de la notation entre une dizaine d'examinateurs ou plus. On sait que certains notent beaucoup plus sévèrement que d'autres et qu'un étudiant, évalué par un autre enseignant, lorsqu'on procède à un échange de groupes à l'occasion d'un test oral peut obtenir une note très éloignée de celle qu'il attendait. De gros écarts subsistent, même à l'issue d'une session de formation où l'équipe d'examinateurs s'était entraînée à partir de documents vidéos montrant des étudiants de divers niveaux en situation d'examen.

Des écarts se manifestent également suivant que l'évaluateur est de langue maternelle anglaise ou non. Habituellement, les francophones ont tendance à noter plus sévèrement que leurs collègues anglophones - qui ont une plus grande tolérance des erreurs, à condition qu'elles n'entravent pas la compréhension.

Recommandations

Afin de corriger les écarts, il semble essentiel d'établir des critères.

Ceux-ci peuvent être traditionnels : grammaire, vocabulaire, prononciation, fluidité et contenu. Ils peuvent aussi inclure les notions de registre et prendre en compte l'interaction.

Les critères serviront de guide à l'examinateur pendant l'épreuve, mais il n'est pas souhaitable d'affecter rigidement d'une note chacun des critères.

Il faut, tout en tenant compte des critères sélectionnés, attribuer la note finale sur une impression d'ensemble, ce qu'Underhill appelle *impression making*. Cette impression ne peut échapper à la subjectivité. Mais nous sommes toujours évalués subjectivement lorsque nous participons à une interaction, que ce soit dans notre langue maternelle ou en langue étrangère.

Françoise Cormon, méthodologue chevronnée aux Etudes Pédagogiques de Genève, confirme ce point de vue. Après avoir passé des années à peaufiner des grilles d'évaluation pour l'oral, elle les délaisse à présent en faveur d'un jugement global basé sur une liste de critères. Elle a observé que les écarts de notation entre examinateurs, à partir d'une impression d'ensemble par rapport à un niveau moyen requis, étaient moindres ainsi que lorsque les enseignants effectuaient un total de points pour tous les critères.

Conclusions

En dépit des résistances rencontrées (difficultés de mise en place, perte de temps, manque de fiabilité) un examen oral est la seule façon d'évaluer l'oral - si le développement des capacités à s'exprimer à l'oral est l'objectif d'un cours de langue. La notation par deux examinateurs - qui entraîne des frais supplémentaires - semble un moyen très sûr de pallier les disparités d'évaluation mentionnées ci-dessus.

Références bibliographiques

- Underhill, N. 1987. ***Testing Spoken Language***. Cambridge: Cambridge University Press.

Ce livre offre une description très complète des différents types de tests. Son but est d'encourager les examinateurs à expérimenter diverses techniques et à en changer si elles ne conviennent pas, plutôt que de suivre automatiquement une façon de faire. Les opérations statistiques pour assurer la fiabilité des notes d'un même examen et entre plusieurs épreuves conviennent pour des épreuves écrites mais ne sont pas adaptées à l'oral, par essence plus fluide et spontané. L'auteur insiste sur l'aspect humain de la rencontre entre deux personnes que constitue un test oral - à distinguer de "*la machine à saucisses*" des résultats statistiques. Il faut rétablir l'équilibre entre traitement statistique et évaluation subjective.

- Hughes, A. 1989. ***Testing for Language Teachers***. Cambridge: Cambridge University Press.

Dans ce livre plus complet, le test oral n'est qu'un chapitre parmi les autres sortes de tests, pour l'écrit, la lecture, l'écoute, la grammaire et le vocabulaire. L'auteur insiste sur la nécessité de cohérence entre test et enseignement. On trouve un chapitre sur la *fiabilité* (les résultats obtenus avec un groupe de niveau donné sont similaires à ceux qu'on obtiendra à une autre occasion), un autre sur la *validité* (un test mesure avec précision ce qu'il est destiné à mesurer) et une annexe sur l'analyse statistique des résultats de tests.

Anne Péchou

Bibliographie 2

- Alderson, J. Charles & North, Brian. ***Language Testing in the 1990's: The Communication Legacy***. London: MEP/British Council, 1991.
- Allen, J.P.B. & Davies, Alan, eds. ***Testing and Experimental Methods***. Oxford University Press, 1977.
- Baker, David. ***Language Testing: A Critical Survey and Practical Guide***. London: Edward Arnold, 1989.
- Carroll, Brendan J. ***Testing Communicative Performance***. ***An Interim Study***. Oxford: Pergamon Press, 1980.
- Cohen, Andrew D. ***Testing Language Ability in the Classroom***. Rowley, Mass.: Newbury House, 1980.
- Davies, Alan. ***Principles of Language Testing***. Oxford: Blackwell, 1990.
- Davies, Susan & West Richard. ***English Language Examinations***. London: Longman, 1989.
- Fulcher, Glenn. " Tests of Oral Performance: The Need for Data-based Criteria ". *ELT Journal* 41 : 4, October 1987. pp. 287-291.
- Harris, David P. ***Testing English as a Second Language***. New York: McGraw Hill, 1969.
- Harrison, Andrew. ***A Language Testing Handbook***. London: MEP, 1983.
- Heaton, J.B. ***Classroom Testing***. London: Longman, 1990.
- Horner, David. " Testing: Introduction and Review ". *TESOL News* 10 : 1, Spring 1990. Reprinted in *The Best of TESOL France News* 1 : 1, 1994, pp. 109-114.
- Hughes, Arthur. ***Testing for Language Teachers***. Cambridge University Press, 1989. (*contient 3 pages de bibliographie*)
- ***Issues in Language Testing***. ELT Document n° 111, British Council, 1981.
- Knight, Ben. "Assessing Speaking Skills: A Workshop for Teacher Development". *ELT Journal* 46 : 3, July 1992, pp. 294-302.
- *Language Testing* (revue bi-annuelle). London: Edward Arnold.
- Rhea-Dickins, Pauline & Germaine, Kevin. ***Evaluation***. Oxford University Press. 1992.
- Skehan, Peter. " Communicative Language Testing ". *TESOL News* 10 : 1, Spring 1990. Reprinted in *The Best of TESOL France News* 1 : 1, 1994, pp. 115-127.
- " The Testing of Oral Proficiency " *System* 20 : 3 (special issue), August 1992.
- Valette, Rebecca M. ***Modern Language Testing: A Handbook***. New York: Harcourt, Brace & World, 1967.
- Weir, Cyril J. ***Communicative Language Testing***. Prentice-Hall. 1990.

N.D.