IS LANGUAGE PROFICIENCY TESTING DUE FOR A FACELIFT?

(Part II)

TOM C. WHITE

English Language Officer
The British Council

This article is the second of two. In the first (Lenguas Modernas 2, December, 1975) criticisms of currently used objetive Foreign Language (FL) Proficiency test batteries were raised, and a proposal for testing language production was sketched.

The main criticism of objective FL proficiency tests related to their *validity,* and weakness in validity was imputed to faulty test desing. In fact objective tests may have survived for so long because they are highly *reliable,* and this may have tempted test-designers to sacrifice valuable measures of communicative competence simply because they are notoriously difficult to assess reliably. The case most often referred to is the essay. Different examiners or assessors may give widely differing marks to the same script on a typical 'free composition' subject such as 'The Worst Dream I ever Had'. A standard examiners' practice to reduce the unreliability of marker assessment is to average the marks awarded by two independent scorers. This, however, does not go far enough. There are more reliable ways of assessing written performance, and the major defects in this particular case are a) the choice of subject, and b) the type of writing task assigned.

WRITING

i: *the 'Cloze' test*

The nearest approach to a direct assessment of written skills in the FL objective test battery is the 'Cloze procedure'. A reading passage is set, and every n^{th} word is deleted. The task is to fill in the gaps with words which complete the sense of the passage. Obviously there are two alternatives for the scorer —he awards a mark only for correct insertion of the original word deleted and rejects all others, or he accepts sensible alternatives as well. In either case the test is attractive —it is easy to set and score, and the alternatives for each 'choice' are not predetermined as they are in the typical 'key plus distractors' item. The main drawbacks are a) the selection of a reading passage which is a fair task for all candidates (i.e. one which avoids specialized vocabulary or highly idiosyncratic style), b) the fact that this procedure does not reproduce or reflect a typical case of language-in-use, and c) the psychological drawback which might produce under-performance if a candidate is put off by the unfamiliar aspect of a 'mutilated' text and adopts a faulty 'answering strategy'.

In fact the language of the Cloze test is still controlled and pre-selected by the test designer in much the same way as it is in objetive tests of structure and vocabulary. Nevertheless, Cloze tests are becoming increasingly popular and are definitely a step in the right direction if what is required is direct evidence of writing ability.

ii: *The dictation*

Two interesting articles (Oller, 1971, and

Oller and Streiff, 1975) argue that dictation should be reinstated in tests of FL proficiency. The basis of the argument is statistical analysis of a test battery used for placing foreign students at the University of California. Dictation was found to correlate better with the overall result than any other of the five subtests in the battery, and also showed the best inter-subtest correlation. Furthermore, high correlations were established between dictation and Cloze tests. This is seen by Oller and Streiff as evidence that both dictation and Cloze procedure are valid tests of underlying language competence. (It could just as well be argued, however, that both are of consistently *low* validity).

The problem with dictation as a test of FL proficiency lies not so much in the technique, which is very convincingly argued for by Oller, but in the selection of the passage to be dictated and the method of administering the test. The passage is, obviously, a predetermined sequence of sentences, and as such it can be unfair or inappropriate for some candidates and favourable to others in the same way as the Cloze reading/writing test. The possible psychological difficulties arising from the test administrator's intonation, accent, and speed of delivery are great. The errors mentioned by Oller and Streiff show how a candidate can underperform if he adopts the wrong strategy for using the information 'present' in the sound-wave. What is meant here is that, quite early on in the dictation test, a candidate may 'hear' *brand sales* instead of the correct *brain cells*. This initial error causes him to process what he hears subsequently along entirely mistaken lines, although, as Oller and Streiff admit, with great ingenuity. One is tempted to add 'with considerable skill in the use of language'. However this skill will not attract credit within the restrictions of the marking scheme for a dictation, and therefore the test almost certainly gives and invalid result in cases such as this.

Both the Cloze procedure and dictation have given encouraging results in large-scale trials, however, and it is reasonable to suppose that a satisfactory psycholinguistic theory will one day show to what extent and under what conditions these tests can validly measure FL proficiency.

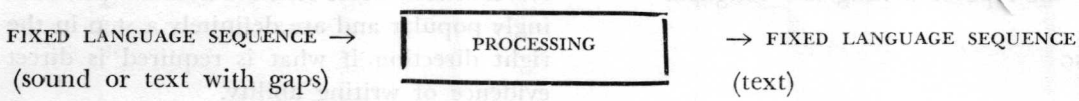
iii: *Controlled-input writing tests*

We now return to the proposal made in the first of these two articles (Lenguas Modernas 2, p. 53). Here are two hypotheses which might prove useful in constructing and evaluating a productive test of writing skills.

a) An independent criterion of the content validiy of a test of writing skill can be established by comparing the activity which the test demands with a description of language-in-use at the required level.

b) An acceptable level of scorer reliability can be achieved by using skilled assessors, trained in consistent evaluation procedures and awarding marks on a uniform and agreed scale.

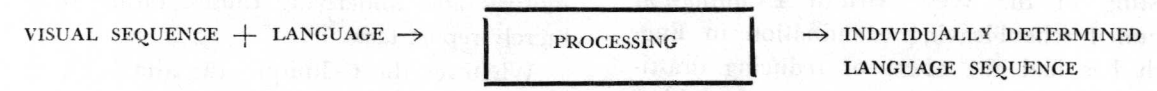
It is with these two hypotheses in mind that the experimental tast of writing skills was devised. It was experimental in that it avoided the exclusive use of language as an "input' or stimulus to the candidate, and sought to maximise the visual element. Dictation and Cloze procedure both restrict the stimulus to arbitrarily-chosen texts, and require the response or 'output' to be the result of satisfactory processing of this stimulus. The sequence is thus.



This is objected to on the grounds that the essential activity (processing) is under the strict control of a predetermined language input. The result of this processing should ideally be an exact reproduction in written mode, of the input. Language-in-use, on the

other hand, is almost never concerned with exact reproduction. The processing is always interpretative and the output is a *modified* (often an enriched) version of the input.

The decision to use a film for the initial stimulus was therefore taken in an attempt to free the candidate as far as possible from this unwanted language control of his processing activity. It was hoped that the output would be the result of the candidate's autonomous, interpretative decision and would provide a rich array of language data for the assessor to evaluate. The sequence would therefore be:



The attractiveness of this procedure is that it preserves the autonomy of the candidate, who himself produces the sample of the FL on which he is to be evaluated (as in the traditional essay test) while controlling the input in a way which ensures that each candidate is exposed to precisely the same stimuli (thus avoiding the greatest defect of the essay test).

In addition, the assessor can easily acquaint himself with all aspects of the stimulus by viewing the film as often as he needs to.

The two types of film were chosen for the following reasons. Film 1 is an action sequence. A girl is walking along the street when she hears someone scream in a house nearby. She stops a passer-by and asks him to help. The passer-by initially disbelieves the girl but then both hear another scream coming from the same house. They stop a man with a ladder and persuade him to put it against the wall of the house near an open window on the upstairs floor. Just as the first man is climbing up the ladder the front door of the house opens, and a woman in a nightdress emerges and angrily orders all three would-be helpers to depart. The candidate's report of what happened tests the following skills:

i ability to report events in sequence

ii ability to use appropriate everyday vocabulary

iii ability to process visual information and a comparatively meagre dialogue component, and transform into a written report.

Satisfactory perfomance in these three tasks indicates the communicative competence of the candidate. The film lasts for nine minutes; the report is to be written in thirty minutes. The candidate is encouraged to take notes (in any language) while watching the film. Basically this test is a language expansion exercise —it involves composing a connected account which makes sense of a connected sequence of events without the original visual element.

The second film has a much fuller soundtrack but a much simpler visual element (cartoons and tables). The task here is to compress the input into a summary. The subject-matter is the Colombo Plan and its aid programme for development in third-world Asian countries. The tasks which need to be accomplished for a satisfactory performance in this test are:

i to select and express the main ideas

ii to subordinate or eliminate secondary data such as exemplification

iii to write up the selected data coherently

It will be seen that neither of these tests is 'pure'. They involve the co-ordination of complex mental skills and as such are at the opposite pole to tests which seek to identify and isolate discrete components of a skill and test them one by one. In this connection it is worth asking whether the elaborate attempts to isolate the elements of language for the purpose of objective testing have any theoretical justification at the level of FL proficiency. As Eastwood (1964) remarks: 'Is the de-

sign appropriate for the data or are the data being made to fit the design?' In testing for FL proficiency we are testing for advanced language skills. These are essentially active, interpretative and integrative and it seems perverse to try to reduce them to component parts. Indeed the introduction of objective testing in the West African Examination Council school-leaving examination in English has had the effect of reducing drastically the classroom time spent on teaching the language and increasing (sometimes up to 75% of total hours) the time spent on working through objective tests. (Forrest, 1975).

Tests of advanced language skills ought to encourage a type of teaching which imparts the mental and cognitive abilities indispensable for adequate performance in the FL.

ASSESSMENT

The approach to productive language testing advocated in these articles requires expert assessment if the test are to be reliable as well as valid. There is therefore no place for the unskilled or mechanical assessment processes which can be applied to objective test batteries. It is relevant to make an appeal to the professionalism of teachers and test administrators who have too easily given up their rightful position in the face of the statisticians, and the computer-programmers. Like any other skilled professional (the microbiologist, the radiologist, the doctor or psychiatrist), the expert teacher knows, by the normal exercise of intuitive judgement, based on wide experience and training, what a performace is worth. Like all expert diagnosticians, he will find it irrelevant (probably impossible) to list exhaustively the various attributes of the totality which he is assessing. But he need not worry that this is in any way 'unscientific'.

In assessing a genuine 'language performance', the scorer will attend to the Gestalt or configurational quality of the script, and will mark according to the criteria of

adequacy of coverage (the *content* of the script)
communicative adequacy (the choice of language)
organization of ideas (discourse features)

He may also add a fourth category for scripts of exceptional merit, awarding credit for the ability to evaluate and judge, to comment on motives and underlying causes, rather than merely report facts.

Whatever the technique, the aim is clear: to place each script in one of four categories:

Clear pass
Borderline pass
Borderline failure
Clear failure

Some assessors prefer to have a detailed marking scheme, others work better using their judgement of the whole performance without classifying marks under separate headings.

Instead of the atomistic marking of objective tests where each item can only be correct or false, there is a sliding scale of communicative adequacy for performance on a productive test, running from the totally adequate to the totally inadequate. This implies quite a different, and much more flexible approach to 'errors', which are counted as more or less serious according to how they interfere with the communicative function of the written text. One final point —this test obviously covers listening comprehension as well as writing ability. A further check on listening comprehension is available, if necessary, in the oral part of the productive test battery.

SPEAKING

The structured interview

Language is used when there is something to express or communicate. The test administrator must elicit appropriate language from the candidate in a face-to-face situation. This can be done in a number of different ways, e.g. using photographs, reading texts or prepared topics of interest. The assessment once again follows the same lines —the perform-

ance of the candidate is judged first on its communicative adequacy, with additional consideration given to clarity of diction, fluency and style. Skill is required in putting the candidate at ease, by selecting topics which will allow him to communicate effectively, and this calls for a sympathetic attitude and a flexible approach. Once more, the four-category assessment is all that is needed in assessing proficiency in the spoken language.

Conclusión

These articles have sketched a radically different approach to the assessment of FL proficiency from the one adopted by exponents of objective testing. A particular objection likely to be raised is 'how does one separate a candidate's language ability from his general mental qualities? Is this not an examination rather than a test?' The writer does not see this as a major difficulty. A FL user may well have an agile mind but, poor knowledge of the FL will impede the full display of his mental faculties. On the other hand, communicative competence cannot exist without the corresponding cognitive ability or intelligence. The productive language test battery is not an examination as it is not based on a syllabus or specific course of study. The purely cognitive abilities required are not likely to impede effective performance. The memory is not overstrained, even in the writing assignment, as the most that is called for is recall of the main events or ideas in a nine-minute film.

The likely benefit of such language tests on FL teaching has already been mentioned, but they also offer a rich possibility for accurate diagnosis of language difficulties. The 'data' on which the assessment is made is language freely chosen and produced by each candidate. As such it is much more likely to provide useful evidence of definable needs for future study among candidates in the 'borderline' categories than the disjointed array of 'errors' appearing on the check-sheet of an objective test.

BIBLIOGRAPHY

An + denotes works referred to directly in the two articles.
The other titles are included for the general background to the views expressed. For a more comprehensive bibliography of objective FL testing the reader is advised to consult the British Council Specialized Bibliography B 8: *Language Testing, with special reference to English as a foreign language,* available at the British Council office Santiago, or from ETIC.

1. Bruner, J. S., Goodnow, J. J. and Austin, G. A., *A Study of Thinking* (New York, Wiley, 1956).

2. Catford, J. C., *Intelligibility (ELT Selections 2,* 141-150, London, Oxford University Press, 1967).

3. Davies, A. (ed), *Language Testing Symposium* (London, Oxford University Press, 1968).

+ 4. Eastwood, G. R., *Selected Aspects of Language Usage in Education* (Australian Journal of Higher Education, Vol. 2, 1, November, 1964).

+ 5. Forrest, R., *Objective Examinations and the Teaching of English* (English Language Teaching Journal xxix, 3 April, 1975).

+ 6. Oller, J. W., *Dictation as a Device for Testing Foreign-Language Proficiency* (English Language Teaching, xxv, 3, June, 1971).

+ 7. Oller, J. W. and Streiff, V., *Dictation: A Test of Grammar-Based Expectancies* (English Language Teaching Journal xxx, 1, October, 1975).

8. Polanyi, M, *Knowing and Being* (London, Routhledge, 1969).

9. Thorndike, R. L. and Hagen, E. P., *Measurement and Evaluation in Psychology and Education* (New York, Wiley, 1955).

10. Etic, British Council: *ELT Documents 75/3 (special* issue on language testing).