IS LANGUAGE PROFICIENCY TESTING DUE FOR A FACELIFT?

TOM C. WHITE

English Language Officer
The British Council

This article argues that 'objective' tests of language proficiency are based on a widely discredited theory of language, an outmoded approach to foreign language (FL) teaching, and an unnecessarily narrow view of 'objectivity'. It offers suggestions, largely untried as yet, for a test-battery composed of productive language tests as an alternative more in keeping with present-day theory and practice.

Twenty years ago structuralism in linguistics was still in its prime. Knowledge of a language was confidently believed to be knowledge of a finite repertoire of structural elements which could be combined and compounded to form sentences or utterances, together with knowledge of smaller or larger ranges of lexical items which could be selected and slotted into appropriate gaps in the structural framework. The audiolingual method, based on the psychological theory of conditioning then in favour, seemed admirably suited to teaching this structural repertoire. It was easy to grasp and practice, little skill in writing materials was required, and the body of 'doctrine' associated with the psychological theory reassured those teachers who needed to know exactly where they were in the syllabus and exactly what they could allow their students to do in the classroom and language laboratory. The stimuli, in the form of correctly constructed sentences, required a predetermined response, and there was an immediate check for the teacher on the correctness of the response. Learning was conveniently reduced to producing conditioned responses on exposure to the appropriate stimuli, and everyone knew how well this had been demonstrated by Pavlov's dogs and Skinner's pigeons. This all seemed very scientific and satisfactory, and the set of correct 'habits' which was taught during the conditioning procedure was confidently expected to be transferred to a set of skills which enabled the student to use the FL correctly.

Above all, the theories and teaching methods offered an ideal language testing format. Tests could be constructed according to well-tried and trusworthy techniques of sampling, and evaluated by standard statistical methods.

All that was needed was an adequate sample of the language at the level of difficulty required, the construction of a large number of individual test-items in multiple-choice format, the administering of the test to an adequate sample of the 'target population' (i.e. all the people who knew or claimed to know the FL at that particular level of difficulty), the analysis of the results item-by-item, the retention of 'good' items and rejection of 'bad' ones, and the assembling of a

number of good items to make up the final version of the test.

For the sake of thoroughness the final version of the test would then be administered to another sample of the target population, the results processed stastistically, and a table or 'standard scores' drawn up, against which the results of any individual taking the test on a subsequent occasion could be measured. Strictly speaking, a standard score on an 'objective' proficiency test is a comparison of an individual's performance with the performances of a large number of similar people (e.g. × performed better than 80% of the people in the target population sample, or more poorly than 20%). It does not mean that × 'scored 80%', or even that × 'passed the test'.

Tests, however, are usually tools for a particular purpose and the question of passing or failing is involved whenever a proficiency test is used for selection. The fairest method is to evaluate the performance of a subset of people who have taken the test and are engaged in an activity similar to the one aspired to by the candidate for selection. Their performance in the test is calculated and compared with their record of achievement in the activity (e.g. post-graduate study at an anglophone university). On this basis, at least in theory, a point in the marking scale can be fixed so as to include all those who studied successfully and exclude all those who failed. But as there are many other factors involved in success or failure in post- graduate study, the impressions of course-tutors must also be taken into account. If the tutor thinks that for a certain candidate or group of candidates the main cause of failure or under-achievement is poor English, then the test evaluator can claim to have a guide to the predictive value of the test.

The above is a very inadequate summary of the way 'objective' language proficiency tests work. The practical advantages of such tests can be summarised as follows:

i they are easy and quick to administer
ii identical test-conditions obtain for all candidates (same questions, same time-limits, same tape for aural tests, etc.)
iii they can be administered and marked by non-specialists
iv the marking scale offers a wide spread between better and poorer candidates
v large numbers of candidates can be tested and papers scored in a short time
vi impartiality is assured and scorer-error virtually eliminated.

Against this, the practical disadvantages should be weighed:

i they are extremely time-consuming to construct and evaluate
ii there is a security problem if large numbers of people are tested
iii if the answers are 'leaked' or reconstructed from memory and made available to other candidates the battery becomes useless
iv inflexible use of the 'cut-off point' can be unfair to borderline performers
v candidates under-perform if they fail to understand the test instructions.

All this would be worth the time and effort but for one fatal drawback. It is the frequent experience of 'consumers' of objective test batteries that they can give a very misleading impresion of a candidate's proficiency. Every year some of the candidates who are selected fail in their studies because of language difficulties, despite having scored well over the 'passmark' in the proficiency test. Designers of objective test batteries are unable to explain what goes wrong in these cases. But if a test of FL proficiency fails to eliminate certain candidates, it could also be eliminating other candidates who deserve to be selected. (It would be difficult to prove that this in fact happens. One would have to run an experiment in which a number of people who failed the test were sent to study

as though they had passed, and their progress observed — obviously an expensive and wasteful procedure).

I suggest that the reason for erratic performances in objective tests of FL proficiency is faulty test design. Let us examine a few subtests commonly employed in objective test batteries.

1. *Vocabulary*

This is usually tested by giving a sentence with a word underlined (usually a noun) and four or five multiple-choice options from which the candidate must select the word nearest in meaning to the word underlined.

Defects of multiple-choice vocabulary tests:

a) a 50-item test of nouns is a very small sample of the nouns in a language: what is the basis used for selection of these items?

b) if the nouns are of Latin origin, the test favours native speakers of Latin languages and penalises speakers of unrelated languages

c) failure to select the synonym does not prove failure to understand the underlined word

d) a correct answer does not prove ability to use the item correctly

e) failure to understand the sentence or options does not prove inability to communicate the idea in alternative ways.

Comment: This type of test reflects the 'word-list' approach to language and language-teaching. It is easy to find out whether an individual recognizes a particular vocabulary item. What objective vocabulary tests cannot do is measure the range of an individual's 'vocabulary-for-use'. This is an essential part of his communicative competence. Objective tests of vocabulary merely indicate ability to recognise arbitrarily selected lists of words.

2. Structure

This is commonly tested by giving a sentence with three or four options for a particular structure within the sentence, only one of which is grammatically correct.

Defects of multiple-choice structure tests:

a) many of the grammatically incorrect options do not seriously hamper the communicative function of the sentence

b) the basis for selecting a particular structure for testing in this way is frequently either an examiner's repertoire of 'common errors' or the test-designer's intuition of what constitutes a 'good item' for a test

c) it is impossible to treat an incorrect response as evidence that a candidate is incapable of adequately understanding or communicating the idea in the sentence in some other way

d) a taxonomy of 'structure-for-use' has yet to be compiled, and therefore the structural inventories sampled for the purpose of constructing tests are not classified according to communicative fuction

e) discourse — structures do not easily fit into typical test formats, yet they are crucial for FL proficiency.

Comment: Objective tests of structure are based on a descriptive taxonomy of structures (or surface-structures) which takes no account of communicative function or 'structure-in-use'. A vital language skill at the level of proficiency is the ability to select appropriate structures when encoding messages. The test format does not allow exploration of competence in selection, but only tests ability to recognise pre-selected structures in narrow contexts. The structure of discourse is not normally amenable to objective testing, as it involves other skills (e.g. ability to organize and process data). Structure in discourse is usually consigned to the comprehension subtest in an FL proficiency battery.

3. Phoneme Discrimination.

This is usually tested by running through a selection of phonemic contrasts, and getting candidates to listen to them and mark them 'same'or 'different'. With a little ingenuity a

set of five-choice items can be constructed in this way using three words for each item. The available choices are 1-2-3 same, 1-2-3 diferent, 1-3 same, 1-2 same, 2-3 same.

Defects of multiple-choice phoneme discrimination tests:

a) phoneme contrasts are tested in isolation– a highly artificial procedure since it is very rarely reflected in normal language use. As such its significance in determining proficiency in language use is dubious

b) failure to discriminate between phonemes in the test does not imply inability to recognise phonemic contrasts when listening to connected speech. Meaning and context provide significant cues to the listener, whether he be a native speaker or not

c) the test design seeks to eliminate meaning and context as 'variables' but other factors can affect performance on the subtest, e.g. memory-load, failure to understand the instructions, losing one's place on the answer page.

The above comments are intended to cast doubt on objective tests as valid assessment procedures for determining an individual's proficiency in a foreign language. They are dubious because they take no account of language in situations which require communicative activity. The whole 'scientific' context of objective test design and evaluation needs reexamining in the light of current theories of language in communication. The scientific terms used often prove to be scientific in appearance only. "Objective", for example, has connotations which suggest impartiality, factuality. In fact an objective test is a 'closed' test - it restricts a candidate's choice to those options which have been pre-ordained by the test constructor. The 'objectivity' of objective tests is thus merely and artefact of the test design, not a guarantee of objective assessment of skills in the FL.

Other scientific terms used by objective test designers (item analysis, standard deviation, confidence limits, reliability, etc.) are taken from the vocabulary of statistics. They carry no guarantee of the scientific soundness of the test as an assessment of FL proficiency, but merely indicate mathematically valid ways of processing the scores obtained when the test is used.

To sum up the position so far, we might say that objective FL tests are closed tests which reduce the complex phenomenon of language to an inventory of discrete items presented in low-context or context-free forms in ways which allow candidate's responses to be marked in binary (right/wrong) fashion. Such tests are highly reliable, but their reliability is achieved at the expense of validity.

The writer is deeply dissatisfied with this method of testing for FL proficiency and suggests that productive language tests should be reexamined as an alternative.

Let us take as our definition of proficiency the ability to speak, understand, read and write the FL in advanced study situations. A realistic testing strategy would be to set up a typical situation requiring communication in the FL for each of these four skills and observe the performance of the candidate.

SPEAKING.

The interviewer sets up a situation which requires the candidate to ask purposeful questions in order to find out something not previously known to him. He then introduces another situation in which the candidate has to impart specific knowledge. The speaker's communication skills are evaluated on the grounds of his success in these communication tasks. For example, the candidate is given some details about a car accident and has to invent and ask a series of questions to find out whether the interviewer was involved. In the second task he is asked to give a series of directions (how to perform a particular task or how to get from point X to point Y in his home town).

LISTENING.

The candidate is told to note down data on a specific point and then listens to a lecturette on tape or spoken by the test - administrator. He either reads from his notes or writes up the data requested.

READING.

A reading task is set by the test administrator, for example: 'pay special attention to the parts of the following text which tell you what the author *dislikes*'. A fairly long text is given to the candidate, who is instructed to read it at his normal speed. A time - limit (equivalent to say 200 words per minute) is fixed, and written questions which test the candidate's ability to make a mental summary of what he has read are set.

WRITING.

Two short films are needed for this test. The candidate is shown the first film which has a definite sequence of actions but little or no spoken sound track. He is asked to write down what happened. The second film contains a good deal of expository commentary and the visuals are simple animated cartoons and diagrams. This time, the candidate writes down what the film is about.

The two main academic writing skills – reporting and summarising – are tested in this way.

These tests will be described in more detail in a subsequent article. A great deal remains to be said about scorer reliability and scoring procedures for productive language tests. All that has been attempted so far is the initial presentation of a case for reforming current practice in FL proficiency testing. It can be objected that one major drawback in the alternative proposal is the need for skilled assessment of the productive language test. The writer believes that this is not so great an obstacle as might appear at first sight.