Testing: Introduction and Review

Spring 1990
Volume 10
Number 1

David Horner

Preamble

Before writing *an* article *on* testing, one ought, I suppose, to give one's credentials. Briefly, then, I am quite definitely *not* a testing specialist, but rather a teacher *involved* in testing as part of his job. Concretely, this means that I am a Cambridge oral examiner; that I prepare students for Cambridge First Certificate and Proficiency, Oxford Preliminary and Higher, Franco-British Chamber of Commerce Business English Diploma and The British Institute/University of London business examinations. I am also a member of the examination boards of the Institute business examinations and thus actively engaged in writing papers of both exams.

Test Types

Proficiency tests - i.e. tests designed to assess the *overall* ability of a learner in a foreign language - are one type of assessment. In many ways they are the best known, represented, as they are by big institutional names— TOEFL, Cambridge, etc, and the most sought after in so far as they give a distinct idea of the level attained by any given candidate.

Proficiency tests might be thought of as representing the peak of assessment types. One could then consider that the next step down, so to speak, would be the *achievement test*, i.e. a test which seeks to measure what has been learnt during a course of study. Both *progress achievement tests*, testing interim progress during a course, and *final achievement tests* exist. Although both *can* resemble a proficiency test, the difference lies in, as their name implies what is tested. An achievement test will not take as its basis some hypothetical view of the nature of language proficiency (although such a view may well underly the course), but the course objectives, which will be tested in their entirety or, where this is impossible, sampled.

Moving down another step we come to the *placement test*. Once again, for better or worse, this is one with which we are all familiar. Placement tests can be more or

less well constructed (usually the latter), but as long as the same test is used consistently should give reasonably acceptable results.

Where the problem lies, it seems to me, is in the confusion between achievement tests and placement tests that is common in language teaching institutions and among teachers everywhere. By this I don't, of course, mean that teachers and institutions actually mistake one for the other. I'm thinking, rather, of the promotion of learners from one class to a higher level, after they have (needless to say) demonstrated their mastery of the course objectives and subsequent discovery that new students testing in to that same level directly from the placement test are generally better. This has been a problem whenever I have worked in an institution that operated such a system.

It derives, surely, from the fact that the initial placement test takes no account of the subsequent course objectives for each level (where they exist) and is much more in the nature of a proficiency test sampling a whole range of language. And proficiency tests are notoriously bad indicators of achievement because of the time scale necessary for even a "successful" learner to see quantifiable gains in overall proficiency.

Placement and achievement tests are test types which teachers are commonly asked to write. Yet testing is a specialised field in which it is notoriously difficult to do well. Good tests are rare. Fortunately, two books have appeared recently which offer a great deal of help to teacher-testwriters. The first of these is the excellent CUP publication (1989) by Arthur Hughes, *Testing for Language Teachers*. The other, and happily complementary, is OUP's (1983) *Techniques in Testing* by Harold S. Madsen. The former offers a thoughtful overview of the whole area of testing as well as having chapters on testing the different skills; the latter, an altogether more basic book, is nonetheless a must for the shelves of any institution involved in testing because of its clear, step by step approach to the construction of even the most basic of test items. Hughes is stimulating, Madsen solid, and they make a good pair to have around.

Both placement and achievement tests can also attempt to be *diagnostic tests* - i.e. tests which seek to diagnose which areas a given learner is weak in and requires further or particular study.

At the bottom of the ladder, and potentially the first type a learner might come across, is the *language aptitude test*, i.e. a test which seeks (by test items believed to be necessary sub-skills for the "good" language learner) to identify which candidates will successfully learn a language.

Test Requirements

"Good" tests, it is often said, should be *valid, reliable and economical*. Hughes also makes an excellent case for their importance in encouraging positive *backwash* effects. *Validity* can be of several kinds:

face validity

i.e. does the test *look* "serious" to the candidate? Business English exams, for example, would probably have little face validity if they did not contain business-type tasks, but rather essay questions on the appreciation of English literature.

content validity

i.e. does the test contain the structures, skills, etc. which it is supposed to be concerned with? For example, an achievement test which tested language not covered in the course (and I can still remember a statistics final of this type at university which nearly led to a lynching) would not have content validity.

construct validity

i.e. does the test test *only* the ability under consideration? With *direct* testing, for example of a specific *skill* —testing letter writing by asking candidates to write letters, as in the Oxford exams, for instance —this is fairly straightforward. When one attempts to measure such skills *indirectly*, however—as in the TOEFL there start to be problems, as insufficient is as yet known about what sub-skills actually underly skills in a foreign language. Similarly, a listening test which made considerable demands on reading ability or memory would lack construct validity.

concurrent validity

i.e. do candidates test scores on this particular test compare with their scores on other tests? Someone who got a B on Cambridge Proficiency, for exarnple, should not fail the Oxford Higher or score 300 on the TOEFL.

predictive validity

i.e. do candidates actually perform as well (or as badly) in real life as their test scores would lead one to expect? This is a constant problem for institutions and employers admitting candidates who scored "500 on the TOEFL" or "15 on the Bac". Past experience has shown most good (i.e. choosy) American universities that a TOEFL score of 600 is a reasonably good indicator of language proficiency and that

students can be safely admitted. But even here TOEFL has recently submitted to the obvious and now permits (although it does not yet include) a writing component.

Tests are ***reliable*** if they would consistently give (more or less) the same result, irrespective of when taken or who the markers are. In this respect it is clearly safer to have *objective* marking (e.g. multiple choice questions) where there is only one or a very restricted number of possibilities and the only scorer variable differing abilities to add up. Indeed, to cater for this, most "big" tests now computer score these.

However, objective marking, with discrete point type questions (i.e. one element at a time) is necessarily noncommunicative, which would seem counterproductive, nonsensical and downright silly when we are, as a teaching body, bending over backwards to promote "communication" in our classrooms. Yet, communicative testing is necessarily holistic *(integrated* to use the jargon) and *subjective*. What, therefore, happens to the reliability? Hughes devotes some time to his question, as does Nic Underhill in his highly readable, no-nonsense *Testing Spoken Language*, CUP (1987). Basically, the answer provided by both authors is the same: assessor training. Indeed Underhill gives a suggested training outline for training oral assessors which could just as easily be applied to other areas.

Economy simply implies that a test should be relatively easy to administer at a reasonable cost.

Backwash is a particularly interesting idea in so far as it is generally totally neglected. What is suggested—and its ramifications are everywhere, as demonstrated by Hughes—is that to encourage and achieve good teaching it is more than helpful to have good tests. This is obvious given the pressure which teachers with exam classes are under to "prepare" their learners for the exam in question. Hence, anyone who has had to prepare for TOEFL will well remember the soul-destroying hours of ploughing through MCQs. Strutt emphasises this point in his article. He may well have had (at least the very old version of) the Franco-British Chamber of Commerce and Industry Commercial English examination in mind with its dictation and thème/version, the preparation for which was, alas, in many classes radically removed from the kind of teaching one would like to see going on.

Means

By means I have in mind *how* language can be tested. The articles which follow provide a fair range of possibilities and for those interested in writing their own tests, the two books already mentioned, by Hughes and by Madsen, provide a wealth of essential information, as does Underhill's little gem, despite restricting itself to oral testing. There are two means which are neglected by these authors, however, (with the exception of Underhill) and which I would therefore like to mention: *self-assessment and peer assessment.*

Self-assessment is dealt with convincingly and at length in Brindley. One's first reaction is to question its reliability, yet Brindley cites experimental evidence tending to demonstrate that self-assessment is as effective as traditional tests as far as placement testing is concerned; and, of course, considerably more economical. Coincidentally, an article by Blue in *Individualizalion and Autonomy in Language Learning (ELT Documents 131)* published by MEP and the British Council (1988) suggests there may be socio-cultural limits to what one can expect from self-assessment, as learners from certain backgrounds tend to consistently over-rate themselves, whereas others seem to under-rate themselves. Underhill notes a similar phenomenon.

However, wherever one is interested in encouraging learner autonomy and developing learner training, there would seem to be a place for self-assessment, at least in terms of placement and achievement tests. Clearly, however, wherever results may be a determining factor in, for example, one's school or career development and cheating may therefore be expected, some external test of a more traditional nature will be necessary. This applies equally to *peer assessment* which is discussed in a short article by Lynch in the same book as Blue's article. Lynch quite rightly points out that, given the subjective nature of assessment of the productive skills, it is best to have a variety of opinions on the success of the communication.

Clearly, this needs to be directed and training is essential—as for self-assessment—but Lynch argues that it does give more reliable results.

(continued overleaf)

Bibliography

Blue et Lynch, "Individualisation and Autonomy in Language Learning", (ELT Documents 131) MEP and British Council
Hughes, Arthur, Testing for Language Teachers, CUP (1989)
Madsen, Harold S., Techniques in Testing, OUP (1983)
Underhill, Nick, Testing Spoken Language, CUP (1987)