Longman Handbooks for Language Teachers

Writing English Language Tests

J. B. HEATON

Writing English Language Tests

A Practical Guide for Teachers of English
as a Second or Foreign Language

J. B. Heaton

Longman

Preface

Attitudes towards language testing have been too narrowly confined for too long. Not until recently has language testing been viewed in its proper light as a potentially powerful and progressive force in teaching. Indeed, the whole subject of testing language opens the door not only to a closer analysis of the testing and teaching methods involved but also to a better appreciation of the nature of the language being taught.

This is a practical handbook on the construction of English tests, intended primarily for the teacher in the classroom. In addition to outlining the general principles of language testing, the book shows the teacher how to construct a wide range of tests and test items and how to select those most suitable for his individual purposes. Several types of test items are described and their value assessed in relation to the particular language elements or skills which they are being used to measure. The ability to construct valid and reliable test items, however, is achieved only after constant practice. Consequently, the latter part of the book contains a carefully planned and balanced programme of work, giving the teacher practice in the different processes of constructing a variety of useful test items. It is hoped that in this way the teacher will develop a deeper insight into the fundamentals and techniques of both the testing and teaching of English as a second or foreign language.

Acknowledgements

Achievement/Attainment Tests The terms *achievement* and *attainment* are generally used to refer to more formal tests which have been designed to show mastery of a particular syllabus (e.g. end-of-year tests, school-leaving examinations, public tests). These tests are similar to class progress tests in the sense that they are generally based on a syllabus and measure what has been taught and learnt. However, they are rarely constructed by the classroom teacher for a particular class and they are designed primarily to measure individual performance rather than to act as a means of motivating the student or reinforcing learning.

Proficiency Tests Proficiency tests measure a student's achievement in relation to a specific task which he will later be required to perform. For example, does a student know enough English to follow a particular course given in the medium of English or to do a particular job requiring a use of English? Proficiency tests rarely take into account any syllabus which the student has followed, since they are concerned with future performance rather than past achievement and are often administered to students from various language-learning backgrounds.

Aptitude Tests Aptitude tests measure the student's *probable* performance in learning a foreign language, showing whether the student has any special aptitude for learning a new language. Will he experience difficulty in identifying the sounds of the new language or its grammatical structures?

Diagnostic Tests Many progress, achievement and proficiency tests can be used as diagnostic tests to some degree, enabling the teacher to identify specific areas of weakness and difficulty so that he is able to plan the most appropriate remedial programme. A diagnostic test is primarily designed to assess the student's knowledge and skills in particular areas before a course of study is begun.

Contents

9 Criteria and types of tests

10 Interpreting test scores — 167

Appendix — 185

Selected bibliography — 189

Practical work — 195

(Practice sections)

1 Introduction to Language Testing

1.1 Testing and teaching

It is unfortunate that so many examinations in the past have led to a separation of testing from teaching. Both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other. Tests may be constructed primarily as devices to reinforce learning and to motivate the student, or primarily as a means of assessing the student's performance in the language. In the former case, the test is geared to the teaching that has taken place, whereas in the latter case the teaching is often geared largely to the test. Standardised tests and public examinations, in fact, generally exert such a considerable influence on the average teacher that they are often instrumental in determining the kind of teaching that takes place before the test.

Consider the effect of the following type of test item on the teaching of English:

> Analyse into clauses, stating the kind and function of each clause:
> We cannot go until we have finished these exercises.

when compared with the effect of the following item:

> Rewrite each of the following sentences in another way but do not change the meaning. Begin each new sentence with the words given.
> We cannot go until we have finished these exercises.
> When . . . .

The former test item encourages teaching *about* the language while the latter encourages practice in using the language.

It can be argued with some justification that language examinations in the past have exerted a harmful influence on the language teacher and have considerably inhibited language learning by encouraging teachers to teach *about* the language. As a consequence, relatively few teachers sought to provide the maximum opportunity for their students to practise the language itself.

Fortunately, many external examining bodies today seek to measure the candidate's actual performance in the language, and in this way sometimes exert a beneficial influence on syllabuses and teaching strategies. Yet, however much concerned a public examining body may be about the

effects on teaching (i.e. the backwash effects) of its particular examination, the main purpose of that examination is to measure the candidate's ability to use the target language. The examination is thus primarily an instrument for measuring and evaluating performance.

1.2   Why test?

The function indicated in the preceding paragraph provides one of the answers to the question: Why test? But it must be emphasised that it is only one of the functions of a test and that furthermore, as far as the practising teacher is concerned, it is perhaps one of the more negative functions.

Although most teachers also wish to evaluate individual performance, the aim of the classroom test is different to that of the external examination. While the latter is generally concerned with evaluation for the purpose of selection, the classroom test is concerned with evaluation for the purpose of enabling the teacher to increase his own effectiveness by making adjustments in his teaching to enable certain groups of students or individuals in the class to benefit more. Too many teachers gear their teaching towards an ill-defined 'average' group without taking into account the abilities of those students in the class who are at either end of the scale.

A good classroom test will also help to locate the precise areas of difficulty encountered by the class or by the individual student. Just as it is necessary for the doctor first to diagnose his patient's illness, so it is equally necessary for the teacher to diagnose his student's weaknesses and difficulties. Unless the teacher is able to identify and analyse the errors a student makes in handling the target language, he will be in no position to render any assistance at all through appropriate anticipation, remedial work and additional practice.

The test should also enable the teacher to ascertain which parts of the language programme have been found difficult by the class. In this way, the teacher can evaluate the effectiveness of the syllabus as well as the methods and materials he is using. The test results may indicate, for example, certain areas of the language syllabus which have not taken sufficient account of L1 learner difficulties or which, for some reason, have been glossed over.

A test which sets out to measure a student's performance as fairly as possible without in any way setting traps for him can be effectively used to motivate the student. A well-constructed classroom test will provide the student with an opportunity to show his ability to recognise and produce correct forms of the language. Provided that details of his performance are given as soon as possible after the test, the student should be able to learn from his errors and consolidate the pattern taught. In this way a good test can be used as a valuable teaching device

1.3   What should be tested and to what standard?

The development of modern linguistic theory has helped to make language teachers and testers aware of the importance of analysing the language being tested. Modern descriptive grammars (though not yet primarily intended for foreign language teaching purposes) are replacing the older, Latin-based prescriptive grammars: linguists are examining the whole complex system of language skills and patterns of linguistic behaviour. Indeed, language skills are so complex and so closely related to many other non-linguistic skills (gestures, eye-movements, etc.) that it may often seem impossible to separate them for the purpose of any kind of assessment. A person always speaks and communicates in a particular situation at a particular time. Without this kind of context, language may lose much of its meaning.

Before a test is constructed, it is important to question the standards which are being set. What standards should be demanded of learners of a foreign language? Should foreign language learners, for example. be expected to communicate with the same ease and fluency as native speakers? Are certain habits of second language learners regarded as mistakes when these same habits would not constitute mistakes when belonging to native speakers? What, indeed, is "correct" English?

Examinations in the written language have in the past set artificial standards even for native speakers and have often demanded skills similar to those acquired by the great English essayists and critics. In imitating first language examinations of written English, however, second language examinations have proved far more unrealistic in their expectations of the performances of foreign learners, who have been required to re-write some of the greatest literary masterpieces in their own words or to write original essays in language beyond their capacity.

1.4   Testing the language skills

Four major skills in communicating through language are often broadly defined as listening, speaking, reading and writing. In many teaching situations it is desirable that areas of the language are first presented orally before reading and writing are practised. Where this is the case, it is important for the test writer to include those types of questions which appear relevant to the ability to speak the language: e.g. questions testing the student's ability to manipulate structures encountered in the spoken language and to make the correct response to a given stimulus.

Success in traditional examinations all too often simply demonstrates that the student has been able to perform well in the examination he has taken and very little else. For example, the precis exercise (one of the important components of many traditional examinations) measures a skill

which is more closely associated with examinations and answering techniques than with language used in real-life situations. In this sense, the traditional examination may tell us relatively little about the student's general fluency and ability to handle the target language, although it may give some indication of the student's ability in some of the skills he needs as a student.

Ways of assessing performance in the four major skills may take the form of tests of:

(1) listening (auditory) comprehension, in which single utterances, dialogues, talks and lectures are given to the testee;
(2) speaking ability, usually in the form of an interview, a picture description, and reading aloud;
(3) reading comprehension, in which questions are set to test the student's understanding of a written text; and
(4) writing ability, usually in the form of essays, letters and reports.

It is the test constructor's task to assess the relative importance of these skills at the various levels and to devise an accurate means of measuring the student's success in developing these skills. Many test writers consider that their purpose can best be achieved if each separate skill can be measured on its own. But it is usually extremely difficult to separate one skill from another, for the very division of the four skills is an artificial one and the concept itself constitutes a vast over-simplification of the issues involved in communication.

1.5   Testing the language elements

In order to isolate the components of the language skills for both teaching and testing purposes, we should look closely at the following elements:

(i)   phonology (concerned with pronunciation, stress and intonation)
(ii)   vocabulary (concerned with word meanings and word arrangements)
(iii) grammar

(i) Tests of phonology

Test items designed to test phonology might attempt to assess the following sub-skills: ability to recognise and pronounce the significant sound contrasts of a language, ability to recognise and use the stress patterns of a language, and ability to hear and produce the melody or the patterns of the tunes of a language (i.e. the rise and fall of the voice).

The following is an example of a test item designed to test phoneme discrimination. Three words are spoken by the teacher (recorded on tape in the case of many public examinations) and the testee is asked to indicate which words are the same. Sometimes, all three words[1] are the same

(AAA); sometimes only the first and second words are the same (AAB) and so on.

(Spoken)			(Written)				
dock	dock	dock	AAA	AAB	ABA	ABC	O
bad	bad	bat	AAA	AAB	ABA	ABC	O
ship	sheep	ship	AAA	AAB	ABA	ABC	O
fairy	furry	ferry	AAA	AAB	ABA	ABC	O

(ii) Tests of vocabulary

A test of vocabulary measures the student's knowledge of the meaning of certain words and word groups. Such a test may test the student's active vocabulary (the words he should be able to use in speaking and in writing) or his passive vocabulary (the words he should be able to recognise and understand when he is listening to someone or when he is reading). Obviously, in this kind of test the method used to select the vocabulary items (= sampling) is of the utmost importance.

> [2] Put a circle round the letter at the side of the word which best completes the sentence.
>
> Did you . . . that book from the school library?
>
> A. beg B. borrow C. hire D. lend E. ask

(iii) Tests of grammar

These tests measure the student's ability to manipulate structures and to distinguish appropriate grammatical forms from inappropriate ones.

> [2] Put a circle round the letter of the word or group of words which best completes each sentence
>
>   I'll wait here until it   . dark.
>
> A. will be
> B. be
> C. is
> D. is being
> E. has been

Note that all three examples in this section are called *multiple-choice items.* The term *multiple-choice* is used because the testee is required to select the correct answer from a choice of several answers. (Only one answer is correct for each item.) The word *item* is used in preference to the word *question* because the latter word suggests the interrogative form; many test items are, in fact, written in the form of statements.

1.6   Language skills and language elements

Items designed to test areas of phonology, vocabulary and grammar will be examined in detail later in the appropriate chapters. The question now

posed is: to what extent should we concentrate on testing the student's ability to handle these elements of the language and to what extent should we concentrate on testing the integrated skills? Our attitude towards this question must depend on both the level and the purpose of the test. If the testee has been learning English for only a relatively brief period, it is highly likely that we shall be chiefly concerned with his ability to handle the language elements correctly. Moreover, if the aim of the test is to sample as wide a field as possible, a battery of tests of the language elements will be useful not only in providing a wide coverage of this ability but also in locating particular problem areas. Tests designed to assess mastery of the language elements enable the test writer to determine exactly what is being tested and to pre-test items, carrying out statistical analyses to determine the degree of reliability and later revising them where necessary.

However, at all levels but the most elementary, it is generally advisable to include test items which measure the ability to communicate in the target language. How important, for example, is the ability to discriminate between the phonemes /iː/ and /i/? Even if they are confused by the testee and he says: *Look at that sheep sailing slowly out of the harbour*, it is unlikely that misunderstanding will result because the context, itself, provides other clues to the meaning. All languages contain numerous so-called "redundancies" which help to overcome problems of this nature. Furthermore, no student can be described as being proficient in a language simply because he is able to discriminate between two sounds or because he has mastered a number of structures of the language. Successful communication in situations which simulate real-life is the best test of mastery of a language. It can thus be argued that fluency in English – a person's ability to express facts, ideas, feelings and attitudes clearly and with ease, in speech or in writing, and his ability to understand what he hears or reads - can best be measured by tests which evaluate performance in the language skills. Auditory and reading comprehension tests, oral interviews and essays assess performance in those language skills used in real life.

Too great a concentration on the testing of the language elements may indeed have a harmful effect on the communicative teaching of the language. There is also at present insufficient knowledge about the weighting which ought to be given to specific language elements. How important are articles, for example, in relation to prepositions or pronouns? Such a question cannot be answered until we know more about the degrees of importance of the various elements at the various stages of learning a language.

1.7   Recognition and production

Methods of testing the *recognition* of correct words and forms often take the following form in tests:

Choose the correct answer and write A, B, C or D.
I've been standing here . . . . half an hour.
A. since
B. during
C. while
D. for

This multiple-choice test item tests the student's ability to recognise the correct form: this ability is obviously not quite the same as the ability to produce and use the correct form in real-life situations. However, this type of item has the advantage of being easy to examine statistically.

If the four choices were omitted, the item would come closer to being a test of *production*.

Complete each blank with the correct word.
I've been standing here . . . . half an hour.

The student would then be required to produce the correct answer (= *for*). In many cases, there would only be one possible correct answer, but production items do not always guarantee that the student will deal with the specific matter the examiner had in mind (as most recognition items do). In this particular case the test item is not entirely satisfactory, for the student is completely justified in writing *nearly/almost/over* in the blank. It would not then test his ability to discriminate between *for* with periods of time (e.g. *for half an hour, for two years*) and *since* with points of time (e.g. *since 2.30, since Christmas*).

The following example, taken from a vocabulary test, also illustrates the difference between testing recognition and testing production:

Choose the best word to complete each blank and write A, B, C or D.
(Recognition)
Stand on the . . . . and tell me how much you weigh.
A. measure   B. weights   C. scales   D. balance

(Production)
Stand on the . . . . and tell me how much you weigh.

A good language test may contain either recognition-type test items or production-type test items, or a combination of both. Each type has its unique functions, and these will be treated in detail later.

1.8   Problems of sampling

The actual question of what is to be included in a test is often difficult simply because a mastery of skills is being assessed rather than areas of knowledge. Although the construction of a language test at the end of the first or second year of learning English is relatively easy if we are familiar

with the syllabus covered, the construction of a test at a fairly advanced level where the syllabus is not clearly defined is much more difficult.

The longer the test, the more reliable a measuring instrument it will be (although length, itself, is no guarantee of a good test). Few students would want to spend several hours being tested – and indeed this would be undesirable both for the tester and the testee. But the construction of short tests which function efficiently is often a difficult matter. Sampling now becomes of paramount importance. The test must cover an adequate and representative section of those areas and skills it is desired to test.

If all the students who take the test have followed the same programme, we can simply choose areas from this programme, seeking to maintain a careful balance between tense forms, prepositions, articles, lexical items, etc. Above all, the kind of language to be tested would be the language used in the classroom and in the student's immediate surroundings or the language required for the school or the work for which the student is being assessed.

If the same mother-tongue is shared by all the testees, the task of sampling is made slightly easier even though the testees may have attended different schools or followed different courses. The testees will all experience problems of a similar nature as a result of the interference of their first-language habits. It is not a difficult matter to identify these problem areas and to include a cross-section of them in the test, particularly in those sections of the test concerned with the language elements. The following two examples based on interference of first-language habits will suffice at this stage. The first example concerns the use of the Present Simple for the Present Perfect tense; thus, many students from certain language backgrounds write such sentences as "Television exists only for the last forty or fifty years" instead of "Television has existed only for the last forty or fifty years". A test item based on this problem area might be:

> Write down A, B, C, D or E according to the best alternative needed to complete the sentence.
> Television . . . . only for the last fifty years.
> A.   exists
> B.   was existing
> C.   has existed
> D.   existed
> E.   is existing

The second example has been taken from a test of vocabulary and concerns confusion in the use of *look for;* it is directed chiefly at Arabic and Chinese learners of English. The word *fetched* has been included in the list of choices because there is no distinction in Arabic between the two concepts expressed in English by *fetch* and *look for,* while account has also been taken of the difficulty many Chinese learners experience as a result of the lack of distinction in Mandarin between *look for* and *find.*