3926

Longman Handbooks for Language Teachers

J. B. Heaton

Writing English Language Tests

New Edition

Consultant editors: Jeremy Harmer and Roy Kingsbury

London and New York

.ongman Group UK Limited
.ongman House, Burnt Mill, Harlow,
'ssex CM20 2JE, England
nd Associated Companies throughout the world.

ublished in the United States of America
y Longman Inc., New York

irst published 1975
hird impression 1990

RITISH LIBRARY CATALOGUING IN PUBLICATION DATA
eaton, J. B.
Writing English language tests. – New ed. – (Longman handbooks for language teachers).
1. English language – Study and teaching – Foreign speakers 2. English language – Ability testing
I. Title
428.2'4'076 PE1128.A2

SBN 0-582-00237-0

IBRARY OF CONGRESS CATALOGUING IN PUBLICATION DATA
eaton, J. B. (John Brian)
Writing English language tests.
(Longman handbooks for language teachers)
Bibliography: p.
Includes index.
1. English language – Study and teaching – Foreign speakers. 2. English language – xaminations – Authorship. 3. English language – Ability testing. I. Title. II. Series.
E1128.A2H394 1988 428.076 87-5273

t in Times Roman

oduced by Longman Group (FE) Ltd.
rinted in Hong Kong

lustrated by David Parkins

.CKNOWLEDGEMENTS

'e are grateful to the following for permission to reproduce copyright material:

he author, John Bright and the University of Cambridge Local Examinations Syndicate for n extract from his critique on specimen examination questions; Harper & Row Publishers nc for a table from p. 140 'ESL Composition Profile' from *Teaching ESL Composition* by ane B. Hughey, Deanna R. Wormuth, V. Faye Hartfield and Holly L. Jacobs (Newbury louse) Copyright © 1983 by Newbury House Publishers Inc; the author, Rosalind Hawkins, hief Examiner for UCLES Preliminary English Test and the University of Cambridge Local xaminations Syndicate for extracts from sample test materials; Hong Kong Education epartment for extracts from the Hong Kong English School Certificate Examination 1968 nd the Hong Kong Secondary Schools Entrance Examination 1968; Longman Group UK Ltd or extracts from *Composition Through Pictures* by J. B. Heaton, *Studying in English* by J. B. leaton and *Writing Through Pictures* by J. B. Heaton; The author, Anthony Tucker for an xtract from his article in *The Guardian* 5th September 1969; and the following examination oards for permission to reproduce questions from past examination papers: Joint latriculation Board; North West Regional Examinations Board; The Royal Society of Arts xaminations Board; University of Cambridge Local Examinations Syndicate; University of Oxford Delegacy of Local Examinations and the Arels Examinations Trust.

Contents

1

Introduction to language testing

1.1 Testing and teaching

A large number of examinations in the past have encouraged a tendency to separate 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, may 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.

A language test which seeks to find out what candidates can do with language provides a focus for purposeful, everyday communication activities. Such a test will have a more useful effect on the learning of a particular language than a mechanical test of structure. In the past even good tests of grammar, translation or language manipulation had a negative and even harmful effect on teaching. A good communicative test of language, however, should have a much more positive effect on learning and teaching and should generally result in improved learning habits.

Compare the effect of the following two types of test items on the teaching of English:

1 You will now hear a short talk. Listen carefully and complete the following paragraph by writing one word on each line:

If you go to . . . . . on holiday, you may have to wait a long time at the . . . . . as the porters are on . . . . . However, it will not be as bad as at most . . . . . .
(etc.)

2 You will now hear a short weather and travel report on the radio. Before you listen to the talk, choose one of the places **A**, **B** or **C** and put a cross (X) in the box next to the place you choose.

Place **A** – Southern Spain (by air). ☐
Place **B** – Northern France (by car). ☐
Place **C** – Switzerland (by rail). ☐

Put crosses in the correct boxes below after listening to the programme. Remember to concentrate only on the information appropriate to the place which you have chosen.

No travel problems	
A few travel problems	
Serious travel problems	
Sunny	
Fine but cloudy	
Rain	
Good hotels	
Average hotels	
Poor hotels	
(etc.)	

Fortunately, a number of well-known public examining bodies now attempt to measure the candidates' success in performing purposeful and relevant tasks and their actual ability to communicate in the language. In this sense, such examinations undoubtedly exert a far more beneficial influence on syllabuses and teaching strategies than in the past. However, even the best public examinations are still primarily instruments for measuring each student's performance in comparison with the performance of other students or with certain established norms.

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 the evaluation of student performance for purposes of comparison or selection is only one of the functions of a test. Furthermore, as far as the practising teacher is concerned, it should rarely be either the sole purpose or even the chief purpose of testing in schools.

Although most teachers also wish to evaluate individual performance, the aim of the classroom test is different from 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 teachers to increase their own effectiveness by making adjustments in their 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 the patient's illness, so it is equally necessary for the teacher to diagnose the 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 or she 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 or she is using. The test results may indicate, for example, certain areas of the language syllabus which have not taken sufficient account of foreign learner difficulties or which, for some reason, have been glossed over. In such cases the teacher will be concerned with those problem areas encountered by groups of students rather than by the individual student. If, for example, one or two students in a class of 30 or 40 confuse the present perfect tense with the present simple tense (e.g. 'I already see that film'), the teacher may simply wish to correct the error before moving on to a different area. However, if seven or eight students make this mistake, the teacher will take this problem area into account when planning remedial or further teaching.

A test which sets out to measure students' performances as fairly as possible without in any way setting traps for them can be effectively used to motivate them. A well-constructed classroom test will provide the students with an opportunity to show their ability to perform certain tasks in the language. Provided that details of their performance are given as soon as possible after the test, the students should be able to learn from their weaknesses. 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 the total context in which they are used as well as to many 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? For example, should foreign language learners after a certain number of months or years 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 rewrite 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, listening and speaking, reading and writing. In many situations where English is taught for general purposes, these skills should be carefully integrated and used to perform as many genuinely communicative tasks as possible. Where this is the case, it is important for the test writer to concentrate on those types of test items which appear directly relevant to the ability to use language for real-life communication, especially in oral interaction. Thus, questions which test the ability to understand and respond appropriately to polite requests, advice, instructions, etc. would be preferred to tests of reading aloud or telling stories. In the written section of a test, questions requiring students to write letters, memos, reports and messages would be used in place of many of the more traditional compositions used in the past. In listening and reading tests, questions in which students show their ability to extract specific information of a practical nature would be preferred to questions testing the comprehension of unimportant and irrelevant details. Above all, there would be no rigid distinction drawn between the four different skills as in most traditional tests in the past, a test of reading now being used to provide the basis for a related test of writing or speaking.

Success in traditional tests all too often simply demonstrates that the student has been able to perform well in the test he or she has taken – and very little else. For example, the traditional reading comprehension test (often involving the comprehension of meaningless and irrelevant bits of information) measures a skill which is more closely associated with examinations and answering techniques than with the ability to read or scan in order to extract specific information for a particular purpose. In this sense, the traditional test 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 scholastic ability in some of the skills he or she needs as a student.

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

- listening (auditory) comprehension, in which short utterances, dialogues, talks and lectures are given to the testees;
- speaking ability, usually in the form of an interview, a picture description, role play and a problem-solving task involving pair work or group work;
- reading comprehension, in which questions are set to test the students' ability to understand the gist of a text and to extract key information on specific points in the text; and
- writing ability, usually in the form of letters, reports, memos, messages, instructions, and accounts of past events, etc.

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. Several test writers still 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 language areas

In an attempt to isolate the language areas learnt, a considerable number of tests include sections on:

- grammar and usage,
- vocabulary (concerned with word meanings, word formation and collocations);
- phonology (concerned with phonemes, stress and intonation).

Tests of grammar and usage

These tests measure students' ability to recognise appropriate grammatical forms and to manipulate structures.

> [1]Although it (1) . . . . . quite warm now, (2) . . . . . will change later today. By tomorrow morning, it (3) . . . . . much colder and there may even be a little snow . . . (etc.)
>
> (1) A. seems B. will seem C. seemed D. had seemed
> (2) A. weather B. the weather C. a weather D. some weather
> (3) A. is B. will go to be C. is going to be D. would be (etc.)

Note that this particular type of question is called *a multiple-choice item*. The term *multiple-choice* is used because the students are required to select the correct answer from a choice of several answers. (Only one answer is normally 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.

Not all grammar tests, however, need comprise multiple-choice items. The following completion item illustrates just one of several other types of grammar items frequently used in tests:

> A: . . . . . does Victor Luo . . . . . ?
> B: I think his flat is on the outskirts of Kuala Lumpur.
> (etc.)

Tests of vocabulary

A test of vocabulary measures students' knowledge of the meaning of certain words as well as the patterns and collocations in which they occur. Such a test may test their *active* vocabulary (the words they should be able to use in speaking and in writing) or their *passive* vocabulary (the words they should be able to recognise and understand when they are listening to someone or when they are reading). Obviously, in this kind of test the method used to select the vocabulary items (= sampling) is of the utmost importance.

In the following item students are instructed to circle 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

In another common type of vocabulary test students are given a passage to read and required to replace certain words listed at the end of the passage with their equivalents in the passage.

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 patterns of the tunes of a language (i.e. the rise and fall of the voice).

In the following item, students are required to indicate which of the three sentences they hear are the same:

Spoken:
Just look at that large ship over there.
Just look at that large sheep over there.
Just look at that large ship over there.

Although this item, which used to be popular in certain tests, is now very rarely included as a separate item in public examinations, it is sometimes appropriate for inclusion in a class progress or achievement test at an elementary level. Successful performance in this field, however, should not be regarded as necessarily indicating an ability to speak.

1.6 Language skills and language elements

Items designed to test areas of grammar and vocabulary will be examined in detail later in the appropriate chapters. The question now posed is: to what extent should we concentrate on testing students' 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 students have been learning English for only a relatively brief period, it is highly likely that we shall be chiefly concerned with their 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.

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 /i:/ and /ɪ/? Even if they are confused by a testee and he or she says *Look at that sheep sailing slowly out of the harbour*, it is unlikely that misunderstanding will result because the context 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 or she is able to discriminate between two sounds or 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 the ability to understand what he or she hears and reads – can best be measured by tests which evaluate performance in the language skills. Listening and reading comprehension tests, oral interviews and letter-writing 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 different stages of learning a language.

1.7 Recognition and production

Methods of testing the *recognition* of correct words and forms of language 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 students' 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.

Students 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 students 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 students are completely justified in writing *nearly/almost/over* in the blank. It would not then test their 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 examples also illustrate the difference between testing recognition and testing production. In the first, students are instructed to choose the best reply in List B for each sentence in List A and to write the letter in the space provided. In the second, they have to complete a dialogue.

(i)

List A	List B
1. What's the forecast for tomorrow?	a Soon after lunch, I think.
2. Would you like to go swimming?	b We can take our umbrellas.
3. Where shall we go?	c All afternoon.
4. Fine. What time shall we set off?	d Yes, that's a good idea.
5. How long shall we spend there?	e It'll be quite hot.
6. What shall we do if it rains?	f How about Clearwater Bay?

(ii) Write B's part in the following dialogue.

1. A: What's the forecast for tomorrow?
   B: It'll be quite hot.
2. A: Would you like to go swimming?
   B: . . . . . .

3. A: Where shall we go?
   B: .....

(etc.)

A good language test may contain either recognition-type items or production-type 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 language skills is being assessed rather than areas of knowledge (i.e. content) as in other subjects like geography, physics, etc. 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 testees. 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 learning 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 students' 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 they may have attended different schools or followed different courses. They 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: 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*. Choices D and E might also appear plausible to other students unsure of the correct use of *look for*.

'Here's your book, John. You left it on my desk.'
'Thanks. I've ..... it everywhere.'

A. looked for
B. fetched
C. found
D. attended to
E. watched over

It must be emphasised that items based on contrastive analysis can only be used effectively when the students come from the same language area. If most of them do not share the same first language, the test must be universal by nature and sample a fair cross-section of the language. It will scarcely matter then if students from certain language areas find it easier than others: in actual language-learning situations they may have an advantage simply because their first language happens to be more closely related to English than certain other languages are. Few would wish to deny that, given the same language-learning conditions, French students learning English will experience fewer difficulties than their Chinese counterparts.

Before starting to write any test items, the test constructor should draw up a detailed table of specifications showing aspects of the skills being tested and giving a comprehensive coverage of the specific language elements to be included. A classroom test should be closely related to the ground covered in the class teaching, an attempt being made to relate the different areas covered in the test to the length of time spent on teaching those areas in class. There is a constant danger of concentrating too much on testing those areas and skills which most easily lend themselves to being tested. It may be helpful for the teacher to draw up a rough inventory of those areas (usually grammatical features or functions and notions) which he or she wishes to test, assigning to each one a percentage according to importance. For example, a teacher wishing to construct a test of grammar might start by examining the relative weighting to be given to the various areas in the light of the teaching that has just taken place: say, the contrast between the past continuous and past simple tenses (40 per cent), articles (15 per cent), time prepositions (15 per cent), *wish* and *hope* (10 per cent), concord (10 per cent), the infinitive of purpose (10 per cent).

Another teacher wishing to adopt a more communicative approach to language testing might consider the following specifications in the light of the learning programme: greeting people (5 per cent), introducing oneself (5 per cent), describing places (15 per cent), talking about the future (20 per cent), making suggestions (5 per cent), asking for information (20 per cent), understanding simple instructions (15 per cent), talking about past events (15 per cent). (It must be emphasised that these lists are merely two examples of the kinds of inventories which can be drawn up beforehand and are not intended to represent a particular set of priorities.) In every case, it is important that a test reflects the actual teaching and the course being followed. In other words, if a more traditional, structural approach to language learning has been adopted, the test specifications should closely reflect such a structural approach. If, on the other hand, a communicative approach to language learning has been adopted, the test

specifications should be based on the types of language tasks included in the learning programme. It is clearly unfair to administer a test devised entirely along communicative lines to those students who have followed a course concentrating on the learning of structures and grammar.

1.9 Avoiding traps for the students

A good test should never be constructed in such a way as to trap the students into giving an incorrect answer. When techniques of error analysis are used, the setting of deliberate traps or pitfalls for unwary students should be avoided. Many testers, themselves, are caught out by constructing test items which succeed only in trapping the more able students. Care should be taken to avoid trapping students by including grammatical and vocabulary items which have never been taught.

In the following example, students have to select the correct answer (C), but the whole item is constructed so as to trap them into making choice B or D. When this item actually appeared in a test, it was found that the more proficient students, in fact, chose B and D, as they had developed the correct habit of associating the tense forms *have seen* and *have been seeing* with *since* and *for*. They had not been taught the complete pattern (as used in this sentence). Several of the less proficient students, who had not learnt to associate the perfect tense forms with *since* and *for*, chose the 'correct' answer.

When I met Tim yesterday, it was the first time I . . . . . him since Christmas.

A. saw
B. have seen
C. had seen
D. have been seeing

Similarly, the following item trapped the more proficient students in a group by encouraging them to consider the correct answer, 'safety', as too simple to be right. Many of these students selected the response 'saturation' since they knew vaguely that this word was concerned with immersion in water. The less proficient students, on the other hand, simply chose 'safety' without further thought.

The animals tried to find . . . . . from the fire by running into the lake.

A. sanitation
B. safety
C. saturation
D. salutation

To summarise, all tests should be constructed primarily with the intention of finding out what students know – not of trapping them. By attempting to construct effective language tests, the teacher can gain a deeper insight into the language he or she is testing and the language-learning processes involved.

Notes and references

Multiple-choice items of this nature have long been used in the United States by such well-known testing organisations as TOEFL (*Test of English as a Foreign Language*, Educational Testing Service, Princeton, New Jersey) and the *Michigan Test of English Language Proficiency* (University of Michigan, Ann Arbor, Michigan) to test grammar and vocabulary. Multiple-choice items have also been widely used in modern language testing in Britain and elsewhere throughout the world. Robert Lado (*Language Testing*, Longman 1961, 1964) was one of the first to develop the multiple-choice technique in testing the spoken language.

2

Approaches to language testing

2.1 Background

Language tests can be roughly classified according to four main approaches to testing: (i) the essay-translation approach; (ii) the structuralist approach; (iii) the integrative approach; and (iv) the communicative approach. Although these approaches are listed here in chronological order, they should not be regarded as being strictly confined to certain periods in the development of language testing. Nor are the four approaches always mutually exclusive. A useful test will generally incorporate features of several of these approaches. Indeed, a test may have certain inherent weaknesses simply because it is limited to one approach, however attractive that approach may appear.

2.2 The essay-translation approach

This approach is commonly referred to as the pre-scientific stage of language testing. No special skill or expertise in testing is required: the subjective judgement of the teacher is considered to be of paramount importance. Tests usually consist of essay writing, translation, and grammatical analysis (often in the form of comments *about* the language being learnt). The tests also have a heavy literary and cultural bias. Public examinations (e.g. secondary school leaving examinations) resulting from the essay-translation approach sometimes have an aural/oral component at the upper intermediate and advanced levels – though this has sometimes been regarded in the past as something additional and in no way an integral part of the syllabus or examination.

2.3 The structuralist approach

This approach is characterised by the view that language learning is chiefly concerned with the systematic acquisition of a set of habits. It draws on the work of structural linguistics, in particular the importance of contrastive analysis and the need to identify and measure the learner's mastery of the separate elements of the target language: phonology, vocabulary and grammar. Such mastery is tested using words and sentences completely divorced from any context on the grounds that a larger sample of language forms can be covered in the test in a comparatively short time. The skills of listening, speaking, reading and writing are also separated from one another as much as possible because it is considered essential to test one thing at a time.

Such features of the structuralist approach are, of course, still valid for certain types of test and for certain purposes. For example, the desire to concentrate on the testees' ability to write by attempting to separate a

composition test from reading (i.e. by making it wholly independent of the ability to read long and complicated instructions or verbal stimuli) is commendable in certain respects. Indeed, there are several features of this approach which merit consideration when constructing any good test.

The psychometric approach to measurement with its emphasis on reliability and objectivity forms an integral part of structuralist testing. Psychometrists have been able to show clearly that such traditional examinations as essay writing are highly subjective and unreliable. As a result, the need for statistical measures of reliability and validity is considered to be of the utmost importance in testing: hence the popularity of the multiple-choice item – a type of item which lends itself admirably to statistical analysis.

At this point, however, the danger of confusing *methods* of testing with *approaches* to testing should be stressed. The issue is not basically a question of multiple-choice testing versus communicative testing. There is still a limited use for multiple-choice items in many communicative tests, especially for reading and listening comprehension purposes. Exactly the same argument can be applied to the use of several other item types.

2.4 The integrative approach

This approach involves the testing of language in context and is thus concerned primarily with meaning and the total communicative effect of discourse. Consequently, integrative tests do not seek to separate language skills into neat divisions in order to improve test reliability: instead, they are often designed to assess the learner's ability to use two or more skills simultaneously. Thus, integrative tests are concerned with a global view of proficiency – an underlying language competence or 'grammar of expectancy'[1], which it is argued every learner possesses regardless of the purpose for which the language is being learnt. Integrative testing involves 'functional language'[2] but not the use of functional language. Integrative tests are best characterised by the use of cloze testing and of dictation. Oral interviews, translation and essay writing are also included in many integrative tests – a point frequently overlooked by those who take too narrow a view of integrative testing.

The principle of cloze testing is based on the Gestalt theory of 'closure' (closing gaps in patterns subconsciously). Thus, cloze tests measure the reader's ability to decode 'interrupted' or 'mutilated' messages by making the most acceptable substitutions from all the contextual clues available. Every *n*th word is deleted in a text (usually every fifth, sixth or seventh word), and students have to complete each gap in the text, using the most appropriate word. The following is an extract from an advanced-level cloze passage in which every seventh word has been deleted:

> The mark assigned to a student . . . . . surrounded by an area of uncertainty . . . . . is the cumulative effect of a . . . . . of sampling errors. One sample of . . . . . student's behaviour is exhibited on one . . . . . occasion in response to one sample . . . . . set by one sample of examiners . . . . . possibly marked by one other. Each . . . . . the sampling errors is almost insignificant . . . . . itself. However, when each sampling error . . . . . added to the others, the total . . . . . of possible sampling errors becomes significant.

The text used for the cloze test should be long enough to allow a reasonable number of deletions – ideally 40 or 50 blanks. The more blanks contained in the text, the more reliable the cloze test will *generally* prove.

There are two methods of scoring a cloze test: one mark may be awarded for each *acceptable* answer or else one mark may be awarded for each *exact* answer. Both methods have been found reliable: some argue that the former method is very little better than the latter and does not really justify the additional work entailed in defining what constitutes an acceptable answer for each item. Nevertheless, it appears a fairer test for the student if any reasonable equivalent is accepted. In addition, no student should be penalised for misspellings unless a word is so badly spelt that it cannot be understood. Grammatical errors, however, should be penalised in those cloze tests which are designed to measure familiarity with the grammar of the language rather than reading.

Where possible, students should be required to fill in each blank in the text itself. This procedure approximates more closely to the real-life tasks involved than any method which requires them to write the deleted items on a separate answer sheet or list. If the text chosen for a cloze test contains a lot of facts or if it concerns a particular subject, some students may be able to make the required completions from their background knowledge without understanding much of the text. Consequently, it is essential in cloze tests (as in other types of reading tests) to draw upon a subject which is neutral in both content and language variety used. Finally, it is always advantageous to provide a 'lead-in': thus no deletions should be made in the first few sentences so that the students have a chance to become familiar with the author's style and approach to the subject of the text.

Cloze procedure as a measure of reading difficulty and reading comprehension will be treated briefly in the relevant section of the chapter on testing reading comprehension. Research studies, however, have shown that performance on cloze tests correlates highly with the listening, writing and speaking abilities. In other words, cloze testing is a good indicator of general linguistic ability, including the ability to use language appropriately according to particular linguistic and situational contexts. It is argued that three types of knowledge are required in order to perform successfully on a cloze test: linguistic knowledge, textual knowledge, and knowledge of the world.[2] As a result of such research findings, cloze tests are now used not only in general achievement and proficiency tests but also in some classroom placement tests and diagnostic tests.

Dictation, another major type of integrative test, was previously regarded solely as a means of measuring students' skills of listening comprehension. Thus, the complex elements involved in tests of dictation were largely overlooked until fairly recently. The integrated skills involved in tests of dictation include auditory discrimination, the auditory memory span, spelling, the recognition of sound segments, a familiarity with the grammatical and lexical patterning of the language, and overall textual comprehension. Unfortunately, however, there is no reliable way of assessing the relative importance of the different abilities required, and each error in the dictation is usually penalised in exactly the same way.

Dictation tests can prove good predictors of global language ability even though some recent research[2] has found that dictation tends to measure lower-order language skills such as straightforward

comprehension rather than the higher-order skills such as inference. The dictation of longer pieces of discourse (i.e. 7 to 10 words at a time) is recommended as being preferable to the dictation of shorter word groups (i.e. three to five words at a time) as in the traditional dictations of the past. Used in this way, dictation involves a dynamic process of analysis by synthesis, drawing on a learner's 'grammar of expectancy'[1] and resulting in the constructive processing of the message heard.

If there is no close relationship between the sounds of a language and the symbols representing them, it may be possible to understand what is being spoken without being able to write it down. However, in English, where there is a fairly close relationship between the sounds and the spelling system, it is sometimes possible to recognise the individual sound elements without fully understanding the meaning of what is spoken. Indeed, some applied linguists and teachers argue that dictation encourages the student to focus his or her attention too much on the individual sounds rather than on the meaning of the text as a whole. Such concentration on single sound segments in itself is sufficient to impair the auditory memory span, thus making it difficult for the students to retain everything they hear.

When dictation is given, it is advisable to read through the whole dictation passage at approaching normal conversational speed first of all. Next, the teacher should begin to dictate (either once or twice) in meaningful units of sufficient length to challenge the student's short-term memory span. (Some teachers mistakenly feel that they can make the dictation easier by reading out the text word by word: this procedure can be extremely harmful and only serves to increase the difficulty of the dictation by obscuring the meaning of each phrase.) Finally, after the dictation, the whole passage is read once more at slightly slower than normal speed.

The following is an example of part of a dictation passage, suitable for use at an intermediate or fairly advanced level. The oblique strokes denote the units which the examiner must observe when dictating.

> Before the second half of the nineteenth century / the tallest blocks of offices / were only three or four storeys high. // As business expanded / and the need for office accommodation grew more and more acute,/ architects began to plan taller buildings. // Wood and iron, however, / were not strong enough materials from which to construct tall buildings.// Furthermore, the invention of steel now made it possible /to construct frames so strong / that they would support the very tallest of buildings. //

Two other types of integrative tests (oral interviews and composition writing) will be treated at length later in this book. The remaining type of integrative test not yet treated is translation. Tests of translation, however, tend to be unreliable because of the complex nature of the various skills involved and the methods of scoring. In too many instances, the unrealistic expectations of examiners result in the setting of highly artificial sentences and literary texts for translation. Students are expected to display an ability to make fine syntactical judgements and appropriate lexical distinctions – an ability which can only be acquired after achieving a high degree of proficiency not only in English and the mother-tongue but also in comparative stylistics and translation methods.

When the total skills of translation are tested, the test writer should endeavour to present a task which is meaningful and relevant to the

situation of the students. Thus, for example, students might be required to write a report in the mother-tongue based on information presented in English. In this case, the test writer should constantly be alert to the complex range of skills being tested. Above all, word-for-word translation of difficult literary extracts should be avoided.

2.5 The communicative approach

The communicative approach to language testing is sometimes linked to the integrative approach. However, although both approaches emphasise the importance of the meaning of utterances rather than their form and structure, there are nevertheless fundamental differences between the two approaches. Communicative tests are concerned primarily (if not totally) with how language is used in communication. Consequently, most aim to incorporate tasks which approximate as closely as possible to those facing the students in real life. Success is judged in terms of the effectiveness of the communication which takes place rather than formal linguistic accuracy. Language 'use'[3] is often emphasised to the exclusion of language 'usage'. 'Use' is concerned with how people actually *use* language for a multitude of different purposes while 'usage' concerns the formal patterns of language (described in prescriptive grammars and lexicons). In practice, however, some tests of a communicative nature include the testing of usage and also assess ability to handle the formal patterns of the target language. Indeed, few supporters of the communicative approach would argue that communicative competence can ever be achieved without a considerable mastery of the grammar of a language.

The attempt to measure different language skills in communicative tests is based on a view of language referred to as the divisibility hypothesis. Communicative testing results in an attempt to obtain different profiles of a learner's performance in the language. The learner may, for example, have a poor ability in using the spoken language in informal conversations but may score quite highly on tests of reading comprehension. In this sense, communicative testing draws heavily on the recent work on aptitude testing (where it has long been claimed that the most successful tests are those which measure separately such relevant skills as the ability to translate news reports, the ability to understand radio broadcasts, or the ability to interpret speech utterances). The score obtained on a communicative test will thus result in several measures of proficiency rather than simply one overall measure. In the following table, for example, the four basic skills are shown (each with six boxes to indicate the different levels of students' performances).

	6	5	4	3	2	1
Listening						
Reading						
Listening and speaking						
Writing						

Such a table would normally be adapted to give different profiles relevant to specific situations or needs. The degree of detail in the various profiles listed will depend largely on the type of test and the purpose for which it is being constructed. The following is an example of one way in which the table could be adapted.

	6	5	4	3	2	1
Listening to specialist subject lectures						
Reading textbooks and journals						
Contributing to seminar discussions						
Writing laboratory reports						
Writing a thesis						

From this approach, a new and interesting view of assessment emerges: namely, that it is possible for a native speaker to score less than a non-native speaker on a test of English for Specific Purposes – say, on a study skills test of Medicine. It is argued that a native speaker's ability to use language for the particular purpose being tested (e.g. English for studying Medicine) may actually be inferior to a foreign learner's ability. This is indeed a most controversial claim as it might be justifiably argued that low scores on such a test are the result of lack of motivation or of knowledge of the subject itself rather than an inferior ability to use English for the particular purpose being tested.

Unlike the separate testing of skills in the structuralist approach, moreover, it is felt in communicative testing that sometimes the assessment of language skills in isolation may have only a very limited relevance to real life. For example, reading would rarely be undertaken solely for its own sake in academic study but rather for subsequent transfer of the information obtained to writing or speaking.

Since language is decontextualised in psychometric-structural tests, it is often a simple matter for the same test to be used globally for any country in the world. Communicative tests, on the other hand, must of necessity reflect the culture of a particular country because of their emphasis on context and the use of authentic materials. Not only should test content be totally relevant for a particular group of testees but the tasks set should relate to real-life situations, usually specific to a particular country or culture. In the oral component of a certain test written in Britain and trialled in Japan, for example, it was found that many students had experienced difficulty when they were instructed to complain about someone smoking. The reason for their difficulty was obvious: Japanese people rarely complain, especially about something they regard as a fairly trivial matter! Although unintended, such cultural bias affects the reliability of the test being administered.

Perhaps the most important criterion for communicative tests is that they should be based on precise and detailed specifications of the needs of the learners for whom they are constructed: hence their particular suitability for the testing of English for specific purposes. However, it would be a mistake to assume that communicative testing is best limited to ESP or even to adult learners with particularly obvious short-term goals. Although they may contain totally different tasks, communicative tests for young learners following general English courses are based on exactly the same principles as those for adult learners intending to enter on highly specialised courses of a professional or academic nature.

Finally, communicative testing has introduced the concept of qualitative modes of assessment in preference to quantitative ones. Language band systems are used to show the learner's levels of

performance in the different skills tested. Detailed statements of each performance level serve to increase the reliability of the scoring by enabling the examiner to make decisions according to carefully drawn-up and well-established criteria. However, an equally important advantage of such an approach lies in the more humanistic attitude it brings to language testing. Each student's performance is evaluated according to his or her degree of success in performing the language tasks set rather than solely in relation to the performances of other students. Qualitative judgements are also superior to quantitative assessments from another point of view. When presented in the form of brief written descriptions, they are of considerable use in familiarising testees and their teachers (or sponsors) with much-needed guidance concerning performance and problem areas. Moreover, such descriptions are now relatively easy for public examining bodies to produce in the form of computer printouts.

The following contents of the preliminary level of a well-known test show how qualitative modes of assessment, descriptions of performance levels, etc. can be incorporated in examination brochures and guides.[5]

WRITTEN ENGLISH

Paper 1 – Among the items to be tested are: writing of formal/informal letters; initiating letters and responding to them; writing connected prose, on topics relevant to any candidate's situation, in the form of messages, notices, signs, postcards, lists, etc.

Paper 2 – Among the items to be tested are: the use of a dictionary; ability to fill in forms; ability to follow instructions, to read for the general meaning of a text, to read in order to select specific information.

SPOKEN ENGLISH

Section 1 – Social English

Candidates must be able to:

(a) Read and write numbers, letters, and common abbreviations.
(b) Participate in short and simple cued conversation, possibly using visual stimuli.
(c) Respond appropriately to everyday situations described in very simple terms.
(d) Answer questions in a directed situation.

Section 2 – Comprehension

Candidates must be able to:

(a) Understand the exact meaning of a simple piece of speech, and indicate this comprehension by:
  – marking a map, plan, or grid;
  – choosing the most appropriate of a set of visuals;
  – stating whether or not, or how, the aural stimulus relates to the visual;
  – answering simple questions.
(b) Understand the basic and essential meaning of a piece of speech too difficult to be understood completely.

Section 3 – Extended Speaking

Candidates will be required to speak for 45–60 seconds in a situation or situations likely to be appropriate in real life for a speaker at this level. This may include explanation, advice, requests, apologies, etc. but will not demand any use of the language in other than mundane and

pressing circumstances. It is assumed at this level that no candidate would speak at length in real life unless it were really necessary, so that, for example, narrative would not be expected except in the context of something like an explanation or apology.

After listing these contents, the test handbook then describes briefly what a successful candidate should be able to do both in the written and spoken language.

The following specifications and format are taken from another widely used communicative test of English and illustrate the operations, text types and formats which form the basis of the test. For purposes of comparison, the examples included here are confined to basic level tests of reading and speaking. It must be emphasised, however, that specifications for all four skills are included in the appropriate test handbook, together with other relevant information for potential testees.[6]

TESTS OF READING

Operations – Basic Level

a. Scan text to locate specific information.
b. Search through text to decide whether the whole or part is relevant to an established need.
c. Search through text to establish which part is relevant to an established need.
d. Search through text to evaluate the content in terms of previously received information.

Text Types and Topics – Basic Level

Form	Type
Leaflet	Announcement
Guide	Description
Advertisement	Narration
Letter	Comment
Postcard	Anecdote/Joke
Form	Report/Summary
Set of instructions	
Diary entry	
Timetable	
Map/Plan	

Format

a. One paper of 1 hour. In addition, candidates are allowed ten minutes before the start of the examination to familiarise themselves with the contents of the source material. The question paper must not be looked at during this time.
b. Candidates will be provided with source material in the form of authentic booklets, brochures, etc. This material may be the same at all levels.
c. Questions will be of the following forms:
   i) Multiple choice
   ii) True-False
   iii) Write-in (single word or phrase)
d. Monolingual or bilingual dictionaries may be used freely.

TEST OF ORAL INTERACTION

Operations – Basic Level

Expressing:	thanks requirements opinions comment attitude confirmation apology want/need information
Narrating:	sequence of events
Eliciting:	information directions service (and all areas above)

Types of Text

At all levels candidates may be expected to take part in dialogue and multi-participant interactions.
The interactions will normally be of a face-to-face nature but telephone conversations are not excluded.
The candidate may be asked to take part in a simulation of any interaction derived from the list of general areas of language use.
However, he will not be asked to assume specialised or fantasy roles.

Format

The format will be the same at each level.

a. Tests are divided into three parts. Each part is observed by an assessor nominated by the Board. The assessor evaluates and score the candidate's performance but takes no part in the conduct of the test.
b. Part I consists of an interaction between the candidate and an interlocutor who will normally be a representative of the school or centres where the test is held and will normally be known to the candidate. This interaction will normally be face-to-face but telepho- formats are not excluded. Time approximately 5 minutes.
c. Part II consists of an interaction between candidates in pairs (or exceptionally in threes or with one of the pair a non-examination candidate). Again this will normally be face-to-face but telephone formats are not excluded. Time approximately 5 minutes.
d. Part III consists of a report from the candidates to the interlocutor (who has been absent from the room) of the interaction from Part II. Time approximately 5 minutes.

As pointed out at the beginning of this chapter, a good test will frequently combine features of the communicative approach, the integrative approach and even the structuralist approach – depending on the particular purpose of the test and also on the various test constraints. If, for instance, the primary purpose of the test is for general placement purposes and there is very little time available for its administration, it m. be necessary to administer simply a 50-item cloze test.

Language testing constantly involves making compromises between what is ideal and what is practicable in a certain situation. Nevertheless this should not be used as an excuse for writing and administering poor tests: whatever the constraints of the situation, it is important to maintain ideals and goals, constantly trying to devise a test which is as valid and reliable as possible – and which has a useful backwash effect on the teaching and learning leading to the test.

Notes and references

1 Oller, J W 1972 Dictation as a test of ESL Proficiency. In *Teaching English as a Second Language: A Book of Readings*. McGraw-Hill

2 Cohen, A D 1980 *Testing Language Ability in the Classroom*. Newbury House

3 Widdowson, H G 1978 *Testing Language as Communication*. Oxford University Press

4 Carroll, B J 1978 *An English Language testing service: specifications*. The British Council

5 The Oxford-Arels Examinations in English as a Foreign Language: *Regulations and Syllabuses*

6 Royal Society of Arts: *The Communicative Use of English as a Foreign Language (Specifications and Format)*

3

Objective testing

(with special reference to multiple-choice techniques)

3.1 Subjective and objective testing

Subjective and *objective* are terms used to refer to the scoring of tests. All test items, no matter how they are devised, require candidates to exercise a subjective judgement. In an essay test, for example, candidates must think of what to say and then express their ideas as well as possible; in a multiple-choice test they have to weigh up carefully all the alternatives and select the best one. Furthermore, all tests are constructed subjectively by the tester, who decides which areas of language to test, how to test those particular areas, and what kind of items to use for this purpose. Thus, it is only the *scoring* of a test that can be described as objective. This means that a testee will score the same mark no matter which examiner marks the test.

Since objective tests usually have only one correct answer (or, at least, a limited number of correct answers), they can be scored mechanically. The fact that objective tests can be marked by computer is one important reason for their evident popularity among examining bodies responsible for testing large numbers of candidates.

Objective tests need not be confined to any one particular skill or element. In one or two well-known tests in the past, attempts have even been made to measure writing ability by a series of objective test items. However, certain skills and areas of language may be tested far more effectively by one method than by another. Reading and vocabulary, for example, often lend themselves to objective methods of assessment. Clearly, the ability to write can only be satisfactorily tested by a subjective examination requiring the student to perform a writing task similar to that required in real life. A test of oral fluency might present students with the following stimulus:

> You went to live in Cairo two years ago. Someone asks you how long you have lived there. What would you say?

This item is largely subjective since the response may be whatever students wish to say. Some answers will be better than others, thus perhaps causing a problem in the scoring of the item. How, for instance, ought each of the following answers to be marked?

ANSWER 1: I've been living in Cairo since 1986.
ANSWER 2: I didn't leave Cairo since 1986.
ANSWER 3: I have lived in the Cairo City for above two years.
ANSWER 4: From 1986.

ANSWER 5: I came to live here before 1986 and I still live here.
ANSWER 6: Since 1986 my home is in Cairo.

Although the task itself attempts to simulate to some degree the type of task students might have to perform in real life, it is more difficult to achieve reliability simply because there are so many different degrees of acceptability and ways of scoring all the possible responses. Careful guidelines must be drawn up to achieve consistency in the treatment of the variety of responses which will result.

On the other hand, reliability will not be difficult to achieve in the marking of the following objective item. The question of how valid such an item is, however, may now be of considerable concern. How far do items like this reflect the real use of language in everyday life?

Complete the sentences by putting the best word in each blank.

'Is your home still in Cairo?'
'Yes, I've been living here . . . . . 1986.'

A. for B. on C. in D. at E. since

Language simply does not function in this way in real-life situations. Consequently, the last item tests grammar rather than communication: it is concerned with students' knowledge of forms of language and how language works rather than with their ability to respond appropriately to real questions.

On the whole, objective tests require far more careful preparation than subjective tests. Examiners tend to spend a relatively short time on setting the questions but considerable time on marking. In an objective test the tester spends a great deal of time constructing each test item as carefully as possible, attempting to anticipate the various reactions of the testees at each stage. The effort is rewarded, however, in the ease of the marking.

3.2 Objective tests

Objective tests are frequently criticised on the grounds that they are simpler to answer than subjective tests. Items in an objective test, however, can be made just as easy or as difficult as the test constructor wishes. The fact that objective tests may generally *look* easier is no indication at all that they *are* easier. The constructor of a standardised achievement or proficiency test not only selects and constructs the items carefully but analyses student performance on each item and rewrites the items where necessary so that the final version of his or her test discriminates widely. Setting the pass-mark, or the cutting-off point, may depend on the tester's subjective judgement or on a particular external situation. Objective tests (and, to a smaller degree, subjective tests) can be pre-tested before being administered on a wider basis: i.e. they are given to a small but truly representative sample of the test population and then each item is evaluated in the light of the testees' performance. This procedure enables the test constructor to calculate the approximate degree of difficulty of the test. Standards may then be compared not only between students from different areas or schools but also between students taking the test in different years.

Another criticism is that objective tests of the multiple-choice type encourage guessing. However, four or five alternatives for each item are sufficient to reduce the possibility of guessing. Furthermore, experience

shows that candidates rarely make wild guesses: most base their guesses on partial knowledge.

A much wider sample of grammar, vocabulary and phonology can generally be included in an objective test than in a subjective test. Although the purposive use of language is often sacrificed in an attempt to test students' ability to manipulate language, there are occasions (particularly in class progress tests at certain levels) when good objective tests of grammar, vocabulary and phonology may be useful – provided that such tests are never regarded as measures of the students' ability to communicate in the language. It cannot be emphasised too strongly, however, that test objectivity by itself provides no guarantee that a test is sound and reliable. An objective test will be a very poor test if:

- the test items are poorly written;
- irrelevant areas and skills are emphasised in the test simply because they are 'testable'; and
- it is confined to language-based usage and neglects the communicative skills involved.

It should never be claimed that objective tests can do those tasks which they are not intended to do. As already indicated, they can never test the ability to *communicate* in the target language, nor can they evaluate actual performance. A good classroom test will usually contain both subjective and objective test items.

3.3 Multiple-choice items: general

It is useful at this stage to consider multiple-choice items in some detail, as they are undoubtedly one of the most widely used types of items in objective tests. However, it must be emphasised at the outset that the usefulness of this type of item is limited. Unfortunately, multiple-choice testing has proliferated as a result of attempts to use multiple-choice items to perform tasks for which they were never intended. Moreover, since the multiple-choice item is one of the most difficult and time-consuming types of items to construct, numerous poor multiple-choice tests now abound. Indeed, the length of time required to construct good multiple-choice items could often have been better spent by teachers on other more useful tasks connected with teaching or testing.

The chief criticism of the multiple-choice item, however, is that frequently it does not lend itself to the testing of language as communication. The process involved in the actual selection of one out of four or five options bears little relation to the way language is used in most real-life situations. Appropriate responses to various stimuli in everyday situations are *produced* rather than chosen from several options.

Nevertheless, multiple-choice items can provide a useful means of teaching and testing in various learning situations (particularly at the lower levels) provided that it is always recognised that such items test *knowledge* of grammar, vocabulary, etc. rather than the ability to *use* language. Although they rarely measure communication as such, they can prove useful in measuring students' ability to recognise correct grammatical forms, etc. and to make important discriminations in the target language. In doing this, multiple-choice items can help both student and teacher to identify areas of difficulty.

Furthermore, multiple-choice items offer a useful introduction to the construction of objective tests. Only through an appreciation and mastery of the techniques of multiple-choice item writing is the would-be test

constructor fully able to recognise the limitations imposed by such items and then employ other more appropriate techniques of testing for certain purposes.

The optimum number of alternatives, or options, for each multiple-choice item is five in most public tests. Although a larger number, say seven, would reduce even further the element of chance, it is extremely difficult and often impossible to construct as many as seven good options. Indeed, since it is often very difficult to construct items with even five options, four options are recommended for most classroom tests. Many writers recommend using four options for grammar items, but five for vocabulary and reading.

Before constructing any test items, the test writer must first determine the actual areas to be covered by multiple-choice items and the number of items to be included in the test. The test must be long enough to allow for a reliable assessment of a testee's performance and short enough to be practicable. Too long a test is undesirable because of the administration difficulties often created and because of the mental strain and tension which may be caused among the students taking the test. The number of items included in a test will vary according to the level of difficulty, the nature of the areas being tested, and the purpose of the test. The teacher's own experience will generally determine the length of a test for classroom use, while the length of a public test will be affected by various factors, not least of which will be its reliability measured statistically from the results of the trial test.

Note that context is of the utmost importance in all tests. Decontextualised multiple-choice items can do considerable harm by conveying the impression that language can be learnt and used free of any context. Both linguistic context and situational context are essential in using language. Isolated sentences in a multiple-choice test simply add to the artificiality of the test situation and give rise to ambiguity and confusion. An awareness of the use of language in an appropriate and meaningful way – so essential a part of any kind of communication – then becomes irrelevant in the test. Consequently, it is important to remember that the following multiple-choice items are presented out of context here simply in order to save space and to draw attention to the salient points being made.

The initial part of each multiple-choice item is known as the *stem*; the choices from which the students select their answers are referred to as *options/responses/alternatives*. One option is the *answer, correct option* or *key*, while the other options are *distractors*. The task of a distractor is to distract the majority of poor students (i.e. those who do not know the answer) from the correct option.

Stay here until Mr Short . . . . . you to come. = *stem*

A. told
B. will tell
C. is telling
D. tells

} = *options/responses/alternatives*

= *distractors*

= *answer/correct option/key*

The following general principles should be observed when multiple-choice items are constructed:

1 Each multiple-choice item should have only *one* answer. This answer must be absolutely correct unless the instruction specifies choosing the *best*

option (as in some vocabulary tests). Although this may seem an easy matter, it is sometimes extremely difficult to construct an item having only one correct answer. An example of an item with two answers is:

'I stayed there until John . . . . . .

A. had come C. came
B. would come D. has come

2 Only one feature at a time should be tested: it is usually less confusing for the testees and it helps to reinforce a particular teaching point. Obviously, few would wish to test both grammar and vocabulary at the same time, but sometimes word order and sequence of tenses are tested simultaneously. Such items are called *impure* items:

I never knew where . . . . . .

A. had the boys gone C. have the boys gone
B. the boys have gone D. the boys had gone

(Note that it may sometimes be necessary to construct such impure items at the very elementary levels because of the severely limited number of distractors generally available.)

3 Each option should be grammatically correct when placed in the stem, except of course in the case of specific grammar test items. For example, stems ending with the determiner *a*, followed by options in the form of nouns or noun phrases, sometimes trap the unwary test constructor. In the item below, the correct answer C, when moved up to complete the stem, makes the sentence grammatically incorrect:

Someone who designs houses is a . . . . . .

A. designer B. builder C. architect D. plumber

The item can be easily recast as follows:

Someone who designs houses is . . . . . .

A. a designer B. a builder C. an architect D. a plumber

Stems ending in *are*, *were*, etc. may have the same weaknesses as the following and will require complete rewriting:

The boy's hobbies referred to in the first paragraph of the passage were

A. camping and fishing
B. tennis and golf
C. cycling long distances
D. fishing, rowing and swimming
E. collecting stamps

Any fairly intelligent student would soon be aware that options C and E were obviously not in the tester's mind when first constructing the item above because they are ungrammatical answers. Such a student would, therefore, realise that they had been added later simply as distractors.

Stems ending in prepositions may also create certain difficulties. In the following reading comprehension item, option C can be ruled out immediately:

John soon returned to . . . . . .

A. work B. the prison C. home D. school

4 All multiple-choice items should be at a level appropriate to the proficiency level of the testees. The context, itself, should be at a lower level than the actual problem which the item is testing: a grammar test item should not contain other grammatical features as difficult as the area being tested, and a vocabulary item should not contain more difficult semantic features in the stem than the area being tested.

5 Multiple-choice items should be as brief and as clear as possible (though it is desirable to provide short contexts for grammar items).

6 In many tests, items are arranged in rough order of increasing difficulty. It is generally considered important to have one or two simple items to 'lead in' the testees, especially if they are not too familiar with the kind of test being administered. Nevertheless, areas of language which are trivial and not worth testing should be excluded from the test.

3.4 Multiple-choice items: the stem/the correct option/the distractors

The stem

1 The primary purpose of the stem is to present the problem clearly and concisely. The testee should be able to obtain from the stem a very general idea of the problem and the answer required. At the same time, the stem should not contain extraneous information or irrelevant clues, thereby confusing the problem being tested. Unless students understand the problem being tested, there is no way of knowing whether or not they could have handled the problem correctly. Although the stem should be short, it should convey enough information to indicate the basis on which the correct option should be selected.

2 The stem may take the following forms:

(a) *an incomplete statement*

He accused me of . . . . . lies.
A. speaking B. saying C. telling D. talking

(b) *a complete statement*

Everything we wanted was *to hand.*
A. under control C. well cared for
B. within reach D. being prepared

(c) *a question*

According to the writer, what did Tom immediately do?
A. He ran home. C. He began to shout.
B. He met Bob. D. He phoned the police.

3 The stem should usually contain those words or phrases which would otherwise have to be repeated in each option.

The word 'astronauts' is used in the passage to refer to

A. travellers in an ocean liner
B. travellers in a space-ship
C. travellers in a submarine
D. travellers in a balloon

The stem here should be rewritten so that it reads:

The word 'astronauts' is used in the passage to refer to travellers in

A. an ocean liner C. a submarine
B. a space-ship D. a balloon