3926

Seoul Branch
YDYC Mellon

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

) Longman Group UK Limited 1988
.ll rights reserved; no part of this publication may be reproduced, stored in a retrieval system,
r transmitted in any form or by any means, electronic, mechanical, photocopying, recording,
r otherwise, without the prior written permission of the Publishers.

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 5
1.1 Testing and teaching 5
1.2 Why test? 6
1.3 What should be tested and to what standard? 7
1.4 Testing the language skills 8
1.5 Testing language areas 9
1.6 Language skills and language elements 10
1.7 Recognition and production 11
1.8 Problems of sampling 12
1.9 Avoiding traps for the students 14

2 Approaches to language testing 15
2.1 Background 15
2.2 The essay-translation approach 15
2.3 The structuralist approach 15
2.4 The integrative approach 16
2.5 The communicative approach 19

3 Objective testing 25
3.1 Subjective and objective testing 25
3.2 Objective tests 26
3.3 Multiple-choice items: general 27
3.4 Multiple-choice items: the stem/the correct option/the distractors 30
3.5 Writing the test 33

4 Tests of grammar and usage 34
4.1 Introduction 34
4.2 Multiple-choice grammar items: item types 34
4.3 Constructing multiple-choice items 37
4.4 Constructing error-recognition multiple-choice items 39
4.5 Constructing rearrangement items 41
4.6 Constructing completion items 42
4.7 Constructing transformation items 46
4.8 Constructing items involving the changing of words 48
4.9 Constructing 'broken sentence' items 49
4.10 Constructing pairing and matching items 49
4.11 Constructing combination and addition items 50

5 Testing vocabulary 51
5.1 Selection of items 51
5.2 Multiple-choice items (A) 52
5.3 Multiple-choice items (B) 56
5.4 Sets (associated words) 58
5.5 Matching items 58
5.6 More objective items 60
5.7 Completion items 62

6 Listening comprehension tests 64
6.1 General 64
6.2 Phoneme discrimination tests 65
6.3 Tests of stress and intonation 68
6.4 Statements and dialogues 69
6.5 Testing comprehension through visual materials 71
6.6 Understanding talks and lectures 82

7 Oral production tests 88
7.1 Some difficulties in testing the speaking skills 88
7.2 Reading aloud 89
7.3 Conversational exchanges 90

7.4 Using pictures for assessing oral production 92
7.5 The oral interview 96
7.6 Some other techniques for oral examining 102

8 Testing reading comprehension 105
8.1 The nature of the reading skills 105
8.2 Initial stages of reading: matching tests 107
8.3 Intermediate and advanced stages of reading: matching tests 110
8.4 True/false reading tests 113
8.5 Multiple-choice items (A): short texts 116
8.6 Multiple-choice items (B): longer texts 117
8.7 Completion items 124
8.8 Rearrangement items 129
8.9 Cloze procedure 131
8.10 Open-ended and miscellaneous items 133
8.11 Cursory reading 133

9 Testing the writing skills 135
9.1 The writing skills 135
9.2 Testing composition writing 136
9.3 Setting the composition 138
9.4 Grading the composition 144
9.5 Treatment of written errors 149
9.6 Objective tests: mechanics 150
9.7 Objective tests: style and register 152
9.8 Controlled writing 154

10 Criteria and types of tests 159
10.1 Validity 159
10.2 Reliability 162
10.3 Reliability versus validity 164
10.4 Discrimination 165
10.5 Administration 167
10.6 Test instructions to the candidate 168
10.7 Backwash effects 170
10.8 Types of tests 171

11 Interpreting test scores 174
11.1 Frequency distribution 174
11.2 Measures of central tendency 175
11.3 Measures of dispersion 176
11.4 Item analysis 178
11.5 Moderating 185
11.6 Item cards and banks 185

Selected bibliography 188
Index 191

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

The same principle applies to grammar items. The following item:

I enjoy . . . . . the children playing in the park.

A. looking to C. looking at
B. looking about D. looking on

should be rewritten in this way:

I enjoy looking . . . . . the children playing in the park.

A. to B. about C. at D. on

If, however, one of the errors made by students in their free written work has been the omission of the preposition after *look* (a common error), then it will be necessary to include *look* in the options.

I enjoy . . . . . the children playing in the park.

A. looking on C. looking at
B. looking D. looking to

4 The stem should allow for the number of choices which have been decided upon. This is particularly relevant, for example, when comparisons are involved in reading comprehension. There is no possible fourth option which can be added in the following item:

Tom was . . . . . the other two boys.

A. taller than
B. smaller than
C. as tall as

The correct

For normal purposes of testing, this should be clearly the *correct* or *best* option: thus, it is most important that each item should be checked by another person.

It can be argued that a greater degree of subtlety is sometimes gained by having more than one correct option in each item. The correct answers in the following reading comprehension and grammar items are circled:

According to the writer, Jane wanted a new racquet because

(A.) her old one was damaged slightly
B. she had lost her old one
C. her father had given her some money for one
(D.) Mary had a new racquet
E. Ann often borrowed her old racquet

Who . . . . . you cycle here to see us?

A. ~~ordered~~ ~~B. caused~~ (C.) made ~~D. asked~~ (E.) let

It is very important, however, to avoid confusing the students by having a different number of correct options for each item, and this practice is *not* recommended. Each of the two multiple-choice test items above actually comprises a group of true/false (i.e. right/wrong) items and, therefore, each alternative should be marked in this way: e.g. in the first item, the testee scores 1 mark for circling A, 1 mark for not circling B, 1 mark for not circling C, 1 mark for circling D, and 1 mark for not circling E (total score = 5).

The correct option should be approximately the same length as the distractors. This principle applies especially to vocabulary tests and tests of

reading and listening comprehension, where there is a tendency to make the correct option longer than the distractors simply because it is so often necessary to qualify a statement or word in order to make it absolutely correct. An example of such a 'giveaway' item is:

He began to *choke* while he was eating the fish.

A. die
B. cough and vomit
C. be unable to breathe because of something in the windpipe
D. grow very angry

The distractors

Each distractor, or incorrect option, should be reasonably attractive and plausible. It should *appear* right to any testee who is unsure of the correct option. Items should be constructed in such a way that students obtain the correct option by direct selection rather than by the elimination of obviously incorrect options. Choice D in the following grammar item is much below the level being tested and will be eliminated by testees immediately: their chances of selecting the correct option will then be one in three.

The present tax reforms have benefited . . . . . poor.

A. that B. the C. a D. an

For most purposes, each distractor should be grammatically correct when it stands by itself: otherwise testees will be exposed to incorrect forms. In the above item (and in all grammar items) it is only the wrong choice, and its implied insertion into the stem, which makes a particular pattern ungrammatical. For example, option A is grammatically correct on its own and only becomes incorrect when inserted into the stem.

The following item (which actually appeared in a class progress test of reading comprehension) contains two absurd items:

How did Picard first travel in space?

A. He travelled in a space-ship.
B. He used a large balloon.
C. He went in a submarine.
D. He jumped from a tall building.

Unless a distractor is attractive to the student who is not sure of the correct answer, its inclusion in a test item is superfluous. Plausible distractors are best based on (a) mistakes in the students' own written work, (b) their answers in previous tests, (c) the teacher's experience, and (d) a contrastive analysis between the native and target languages.

Distractors should not be too difficult nor demand a higher proficiency in the language than the correct option. If they are too difficult, they will succeed only in distracting the good student, who will be led into considering the correct option too easy (and a trap). There is a tendency for this to happen, particularly in vocabulary test items.

You need a . . . . . to enter that military airfield.

A. permutation B. perdition C. permit D. perspicuity

Note that capital letters are only used in options which occur at the beginning of a sentence. Compare the following:

Has . . . . . of petrol increased?

A. the price B. price C. a price

..... of petrol has actually fallen.
A. The price B. Price C. A price

3.5 Writing the test

Where multiple-choice items are used, the testees may be required to perform any of the following tasks:

1 Write out the correct option in full in the blank.

He may not come, but we'll get ready in case he ..... does .....

A. will B. does C. is D. may

2 Write only the letter of the correct option in the blank or in a box (which may appear at the side of the question, etc.).

He may not come, but we'll get ready in case he ..... [B]

A. will B. does C. is D. may

3 Put a tick or a cross at the side of the correct option or in a separate box.

He may not come, but we'll get ready in case he .....

A. will A. ☐
B. does B. ☒
C. is C. ☐
D. may D. ☐

4 Underline the correct option.

He may not come, but we'll get ready in case he .....

A. will B. does C. is D. may

5 Put a circle round the letter at the side of the correct option.

He may not come, but we'll get ready in case he .....

A. will (B.) does C. is D. may

Multiple-choice items are rarely optional in a test, for the testees would then spend considerable time in unnecessary reading before choosing the items they wished to answer. Moreover, unless there are good reasons for weighting different items (using the appropriate statistical methods), it is advisable to award equal marks for each item.

The correct option should appear in each position (e.g. A, B, C, D or E) approximately the same number of times in a test or sub-test. This can usually be achieved if it is placed at random in a certain position or if all the options are placed in alphabetical order (i.e. according to the first letter of the first word in each option). However, if the options have a natural order (e.g. figures, dates), it is advisable to keep to this order.

Blackwell started his career as a lawyer in

A. 1921 B. 1925 C. 1926 D. 1932

Note that few tests consist entirely of multiple-choice items. Many well-known tests strike a happy balance between objective items (including multiple-choice items) and subjective items. They test familiarity with the grammatical and lexical components of the language as well as the ability to use the language productively.

4

Tests of grammar and usage

4.1 Introduction

The following are some of the most common types of objective items used to test awareness of the grammatical features of the language. Each type of item will be treated in greater detail in this chapter.

- multiple-choice items
- error-recognition items
- rearrangement items
- completion items
- transformation items
- items involving the changing of words
- 'broken sentence' items
- pairing and matching items
- combination items
- addition items

It should always be remembered that such items as the above test the ability to recognise or produce correct forms of language rather than the ability to *use* language to express meaning, attitude, emotions, etc. Nevertheless, it is essential that students master the grammatical system of the language they are learning. Thus, classroom tests of grammar and usage can play a useful part in a language programme.

4.2 Multiple-choice grammar items: item types

The type of multiple-choice item favoured by many constructors of grammar tests is the incomplete statement type, with a choice of four or five options. This item may be written in any of the following ways:

Type 1 Tom ought not to . . . . . (A. tell B. having told C. be telling D. have told) me your secret, but he did.

Type 2 Tom ought not to . . . . . me your secret, but he did.

A tell
B. having told
C. be telling
D. have told

Type 3 Tom ought not to

A. tell
B. having told
C. be telling
D. have told

me your secret, but he did.

Type 4 Tom ought not to *have told* me your secret, but he did.

A. *No change*
B. tell
C. having told
D. be telling

Item types 2 and 3 are preferable to 1 because the options do not interrupt the flow of meaning in the sentence: these items present the entire sentence so that it can be read at a glance. Unfortunately, type 1 confuses the reader because of the long parenthesis (i.e. the four options occurring between *ought not to* and *me*). Item type 4 shows the correct (or an incorrect) form as part of the sentence in such a way that it cannot be compared on equal terms with the other options: a correct option, for instance, is generally easier to recognise when it appears in the framework of the sentence than as part of a list of distractors.

Another item type appears below, but it is not recommended since it requires the testees to spend time on unnecessary reading. Not only is it uneconomical but it does not present the 'problem' (i.e. the options) as clearly as item type 2 does.

Type 5
A. Tom ought not to tell me your secret, but he did.
B. Tom ought not to having told me your secret, but he did.
C. Tom ought not to be telling me your secret, but he did.
D. Tom ought not to have told me your secret, but he did.

The following method is useful for testing short answers and responses:

Type 6 'Tom ought not to have told anyone the secret.'

A. 'So ought you.'
B. 'Neither ought you.'
C. 'Neither you oughtn't.'
D. 'So oughtn't you.'

Item type 7 requires the students to select the alternative which is true according to the information conveyed in each sentence. Such an item may be included either in a test of reading comprehension or in a test of grammar: a knowledge of the particular syntax is necessary for the understanding of the sentence.

Type 7 'Tom ought not to have told me.

A. Tom did not tell me but he should.
B. Perhaps Tom may not tell me.
C. Tom told me but it was wrong of him.
D. It was necessary for Tom not to tell me.

It may be argued that an understanding of syntactical patterning is just as necessary for the following item:

..... was Robert late last week?'
Three times.'

A. How much
B. How many
C. How often
D. How long

Items which appear in a test of grammar and structure should be made to sound as natural as possible. The following mechanical test item:

This book belongs to Peter. It is . . . . . .

A. Peter's book
B. the book to Peter
C. the book of Peter
D. the book of Peter's

can be rewritten as follows:

This book belongs to Peter, but that is . . . . . .

A. Mary's book
B. the book to Mary
C. the book of Mary
D. the book of Mary's

Note that distractors should generally be correct both in writing and in speech. The following item proved unsuccessful when it was included in a test because many of the more able students selected option D, the reason being that they pronounced *used to* quite correctly as *use to*/ju:stə/:

I . . . . . to go to my uncle's farm every weekend.

A. am used
B. used
C. was used
D. use

Note that the sample items discussed in this section have so far taken the form of short decontextualised items. In practice, however, such items would all form part of a paragraph or series of paragraphs of descriptive, narrative or expository prose. The provision of a detailed context in this way, however, often limits the range of grammatical features being tested. It is usually impossible, for example, to test the future continuous tense in a narrative set in the past (unless direct speech is used). Similarly, a paragraph describing a simple manufacturing process may not provide the test writer with the opportunity to test all the verb forms and tenses he or she may wish to test. This is the price to be paid for including more natural, contextualised test items. On the other hand, the advantage of such items as that on page 37 lies in the interesting and fairly authentic context (i.e. a newspaper article) which contains the items. This is real language used for a particular purpose. Furthermore, the provision of context helps to ensure that there is only one correct option in each case. Short decontextualised sentences can lead to ambiguity as they are usually open to several interpretations when used as stems for multiple-choice items. For example, option D in the following decontextualised item might be correct (as well as option B) if the student happens to know of a medical research establishment which pays volunteers to assist with research and deliberately catch a cold so as to be able to test various cures!

I couldn't take the test last week because I . . . . . a cold.

A. have caught
B. had caught
C. would catch
D. was catching

Much better for testing purposes is the following item. The passage is taken straight from a newspaper article and thus the language is authentic and unaltered in any way. The context provides students with enough background knowledge and details to avoid ambiguity and alternative interpretations, and the newspaper report itself is very interesting. Does it really matter if it will not allow us the opportunity to test every point of grammar which we may want to test? Students taking this test are being given a real feel for the language they are learning.

A long way from home

A 72-year-old Samoan who (1) . . . . . no English at all spent thirteen days (2) . . . . . on buses in the San Francisco area after had become separated (3) . . . . . his family, police said.

(4) . . . . . said that Faaitua Logo, (5) . . . . . moved to the United States two years ago, left his son and daughter-in-law (6) . . . . . a few minutes in a market in San Jose (7) . . . . . something at a nearby stall. When he tried to return to them, he could not remember where they (8) . . . . . for him.

(9) . . . . . first, he began to walk to their home in Palo Alto, 20 kilometres (10) . . . . . , but later he (11) . . . . . on a bus. H changed from bus to bus (12) . . . . . the daytime and slept unde bushes and trees, police said.

(1) A. is speaking B. speaks C. has spoken D. was speakir
(2) A. to ride B. was riding C. ride D. riding
(3) A. with B. from C. by D. off
(4) A. He B. They C. One D. It
(5) A. which B. that C. who D. what
(6) A. in B. for C. since D. at
(7) A. to buy B. for buying C. and buy D. buying
(8) A. waited B. were waiting C. have waited D. wait
(9) A. For B. On C. In D. At
(10) A. far B. from C. near D. away
(11) A. would jump B. jumped C. has jumped D. would h jumped
(12) A. on B. at C. for D. during

4.3 Constructing multiple-choice items

Although it is not always possible to use samples of students' own writte work to provide the basis for the test items, it should not be too difficult for constructors of classroom tests and school achievement tests to take advantage of the types of errors made by students in their free compositions and open-ended answers to questions.

The following extract from a student's letter is used here and in late sections to show how test items can be constructed. The letter was actua written by a student in a country where English is learnt as a foreign language. The errors have not been 'manufactured' for the purpose of illustration, but they do represent errors made by students from only on particular language background. The mistakes, therefore, will not be typical of mistakes made by students from many other language backgrounds and thus the distractors appearing later may be useless for such students.

There is very much time I didn't write you, and now I have a little free time. Winter is behind us and therefore I hope that you wouldn't mind on such a long period between my last letter and this one. You know how is it. Sun is shining, trees become green and it's difficult to stay closed among walls. Sometimes when the weather is sunny I go to w through the park near my lodging and enjoy looking the children playing. You know, the day before yesterday while looking through r window I saw the wet street and people with umbrellas rushing for money and prestige. I suddenly remembered last summer that belor to us. I suppose that you were not angry to me what happened. I thir that it is not good to discuss about passed feelings.

Item 1
Let us ignore the error in the first sentence for the time being and concentrate on the error of tense after *hope*.

Step 1: The first step is to reduce the length of the sentence and to correct the error (and any other errors in the original sentence). Thus,

> I hope that you wouldn't mind on such a long period between my last letter and this one.

becomes I hope you won't mind waiting for so long.

Step 2: Next we write out the sentence, substituting a blank for the area being tested. We write in the correct option and the distractor which the student has provided for us. However, we have to add a sentence because in certain (rare) contexts, *wouldn't* may be correct.

> I hope you . . . . . mind waiting for so long. I promise to reply sooner in future.
>
> A. won't B. wouldn't

Step 3: We now add another two distractors. Again, we go to the written work of our students to provide these distractors. But if we cannot locate any suitable errors without too much difficulty, we use our own experience and knowledge of the target and native languages. Thus, two useful distractors which would also balance the existing two options might be *shouldn't* and *shan't*.

> I hope you . . . . . mind waiting for so long. I promise to reply sooner in future.
>
> A. won't B. wouldn't C. shouldn't D. shan't

It may be argued, however, that *shan't* is acceptable usage amongst certain speakers, thus giving us two correct options instead of one. Though it is highly improbable that people in most areas would use *shan't*, there is a slight shadow of doubt. This is enough to make it desirable to remove *shan't* from our list of options.

Step 4: One suggestion may be that we replace *shan't* with *can't*. If students from a particular language background make such mistakes as *can't mind*, *can't* should be used as a distractor, and possibly *shouldn't* changed to *couldn't*. As can be seen at this early stage, the actual process of item writing is extremely subjective.

> I hope you . . . . . mind waiting for so long. I promise to reply sooner in future.
>
> A. won't B. wouldn't C. couldn't D. can't

An alternative suggestion for a fourth option might be *don't* or *didn't*:

> I hope you don't mind waiting for so long.
> I hope you didn't mind waiting for so long.

Unfortunately, both *don't* and *didn't* are correct. However, in the following context, *didn't* is not acceptable:

> 'How long are you going to be?'
> 'About half an hour. I hope you . . . . . mind waiting for so long.
>
> A. won't B. wouldn't C. shouldn't D. didn't

It may be argued that *didn't* stands out too much. If so – and if it is equally useful to test the use of *don't* (instead of *won't*) after *hope* – the item could be rewritten as:

'How long will you be?'
'About half an hour. I hope you ..... mind waiting for so long.'

A. don't B. wouldn't C. shouldn't D. didn't

Obviously, there are varying degrees of refinement in the construction of multiple-choice items. Furthermore, some items are much more difficult to construct than others. The following two items based on errors in the student's letter are fairly simple to write.

Item 2

Error: ..... and enjoy looking the children playing.

Item: Old Mr Jones enjoys ..... the children playing.

A. looking C. looking on
B. looking at D. looking to

Some test constructors might be tempted to use *for* as a distractor. It can be argued, however, that *looking for* is correct: old Mr Jones might enjoy looking for the children playing (i.e. he might enjoy walking through the park, chatting to his friends, etc. while he is in the process of looking for his grandchildren, who are playing).

Note that the correct option is now in the third position, C. It is important to vary its position. Note also that the word *looking* appears in each option: in some tests the item might appear as follows:

Old Mr Jones enjoys looking ..... the children playing.

A. – B. on C. at D. to

However, when this format includes a dash (–), it is unnatural and not recommended since the insertion of a dash in the stem would not be normal practice in real life.

Item 3

Error: 'I suppose that you were not angry to me.'

Item: I do hope you weren't angry ..... me.

A. to B. with C. on D. about

Note that *at* is also incorrect and may be used as a possible distractor. On the other hand, it may be felt that a number of native English speakers do say *angry at* a person. The decision whether or not to include *at* in the list of incorrect options is again a very subjective one.

4.4 Constructing error-recognition multiple-choice items

The fourth sentence of the letter on page 37 begins *Sun is shining, trees become green and* ..... The error caused by the omission of the article may be tested as follows, using a multiple-choice item:

..... is shining brightly today.

A. Sun B. The sun C. A sun D. Some sun

It may be argued, however, that the choice here is strictly between options A and B at certain levels where students have learned to avoid using 'a' and 'some' with 'sun'. In such instances, one useful device (still using the multiple-choice format) is the error-recognition type of item.

Type 1

Each sentence contains four words or phrases underlined, marked A, B, C and D. Select the underlined word or phrase which is incorrect or unacceptable.

1. I do (A) hope you wouldn't (B) mind waiting (C) for such (D) a long time.
2. I'm worried (A) that you'll (B) be angry (C) to (D) me.
3. I didn't see (A) Bill since he went (B) into (C) hospital last month (D).
4. My car had broken (A) down, so (B) I went (C) there by (D) foot.

Type 2

There is a mistake in grammar in each of the following sentences. Write the letter of that part of the sentence in which it occurs.

1. Sun (A)/is shining (B)/brightly today (C)/, isn't it? (D)
2. Old Mr Jones (A)/enjoys (B)/looking the children (C)/playing in the park. (D)
3. Tony's father (A)/would not let him (B)/to stay out (C)/late at night. (D)
4. Didn't (A)/Susan tell you (B)/she wouldn't mind to come (C)/with us on the picnic? (D)

Item type 2 allows the test writer to test errors caused by omission: e.g. *Sun is shining* and *looking the children*. This type of error cannot be tested by the first item of the error-recognition type. However, there are different ways of correcting many sentences. For example, students may write B or C to denote the incorrect part of the third sentence above, according to which of these correct versions is in their mind:

Tony's father would not permit him to stay out late. (= B)
Tony's father would not let him stay out late. (= C)

For this reason, the test writer is strongly advised to avoid items of the second type.

Sometimes students are given correct sentences together with the incorrect ones: they are then required to write the letter E if the sentence does not contain any error. In practice, this method does not work too well since many students tend to regard every sentence as containing an error. Indeed, another argument against this type of item is that it emphasises the more negative aspects of language learning. It is clearly not sufficient for students simply to recognise sources of error: they ought to be encouraged at all times to concentrate on recognising and producing the correct forms. This argument is supported by many psychologists and teachers who hold that it is undesirable for students to be exposed too much to incorrect forms. On the other hand, this item type is closely related to those skills required when students check, edit or proof-read any report, article, paper or essay they have just written.

4.5 Constructing rearrangement items

Rearrangement items can take several forms, the first of which to consider here will be the multiple-choice type.

The student who wrote the letter in 4.3 obviously experienced considerable difficulty with word order in reported speech, especially after the verbs *know* and *wonder*. Here are two of the errors he made:

'You know how is it. (3rd sentence)
'I wonder did you grow more fatter since summer.'
(later in the same letter)

If we attempt to test the first error by means of an ordinary multiple-choice item, we are faced with the problem of being restricted to only two options: the correct option and the distractor (i.e. the error).

You know how . . . . . .

A. it is B. is it

As the item stands here, we cannot possibly construct other options. It becomes necessary, therefore, to lengthen the original statement to: *You know how warm it is today*. The item would then read:

'Won t I need a coat?'
'Well, you know how . . . . . .

A. warm is it today
B. today it is warm
C. is it warm today
D. warm it is today
E. today is it warm

There seems to be a danger here of confusing the testees by presenting them with the problem in such a way that a certain amount of mental juggling becomes necessary on their part. A preferable item type[1] is the following word-order item:

Complete each sentence by putting the words below it in the right order. Put in the boxes only the letters of the words.

'Won't I need a coat?'
'Well, you know how . . . . . .'

A. it B. today C. warm D. is ☐☐☐☐

I wonder if . . . . . since summer.

A. grown B. you C. fatter D. have ☐☐☐☐

Word order items are useful for testing other structures and features involving inversion:

Everyone's forgotten . . . . . .

A. cup B. he C. which D. used

Not only . . . . . , but he took me to his house.

A. me B. he C. did D. meet

However . . . . . , you'll never pass that test.

A. you B. try C. hard D. may

Leeds United should have won: just think . . . . . .

A. unlucky B. were C. how D. they

I don't know how long . . . . . .

A. going B. Jim C. is D. to be

The order of adjectives and the position of adverbs can be tested in this way, as indeed can several other grammatical areas:

The police are looking for . . . . . .

A. big B. two C. cars D. black

Would you like to read David Brown's . . . . . .

A. short B. new C. story D. exciting

Tom said . . . . . cleaning his car.

A. had B. finished C. he D. just

Only . . . . . been rude to you!

A. ever B. I C. have D. once

Mrs Walker made Ann . . . . . .

A. her new pen B. to C. show D. me

Someone warned Rob . . . . . thieves.

A. for B. to C. out D. look

In many cases it will be useful to change from a multiple-choice item format to a format involving some actual writing. The rearrangement item can be used to test the same features of word order, but the item format becomes a little less artificial. The students are simply required to unscramble sentences and to write out each sentence, putting the words or constituent parts in their correct order:

1. Not only .....
/the examination/very difficult/unfair/was/but/it/was/also

2. It is not advisable .....
/the examination/late/up/the night/to stay/before

3. The best way to prepare .....
.....
/is/yourself/past papers/timed practice/for the paper/to give/in doing

Note that rearrangement items can be used for sentences as well as for words and phrases. When used for this purpose, such items can offer a means of testing an understanding of connectives and reference devices. Students may be required to write out all the sentences in their correct sequence or simply to put the letters of the various sentences in their correct order:

A. Consequently, you should make every effort to complete the paper.
B. Scribble them down as quickly as you can, if necessary.
C. If you find yourself running out of time before you can complete it, however, don't worry about writing your answers neatly.
D. Remember that it is impossible to score marks on questions which you have not attempted.

4.6 Constructing completion items

Carefully constructed completion items are a useful means of testing a student's ability to produce acceptable and appropriate forms of language. They are frequently preferable to multiple-choice items since they measure

production rather than recognition, testing the ability to insert the most appropriate words in selected blanks in sentences. The words selected for omission are grammatical or functional words (e.g. *to*, *it*, *in*, *is*, *the*): content words may be selected in a vocabulary or reading test.

The error *Sun is shining* in the extract from the student's letter in 4.3 illustrates one (minor) difficulty of constructing satisfactory completion items. Although only one answer is possible here, this completion item would have to appear as:

. . . . . Sun is shining today.
or as: . . . . . sun is shining today.

The former item suggests to the testees that no determiner is necessary (since *Sun* is written with a capital letter) while the latter item suggests that a determiner is necessary (because *sun* is written without a capital).

The item can be simply rewritten as a question to overcome this problem:

Is . . . . . sun shining today?

Here are two more examples of completion items based on the student's letter:

Write the correct word in each blank.

1. The old man enjoys looking . . . . . the children playing.
2. That car belongs . . . . . Helen's mother.
3. I hope you're not angry . . . . . me.

Put *a*, *the*, or *some* in each blank only where necessary. If you think that no word should be placed in the blank, put a cross (×) there.

1. Can you see . . . . . sun shining through the clouds?
2. I saw your uncle . . . . . day before yesterday.
3. What have you been doing since I saw you . . . . . last summer?

Completion items cannot, of course, be machine-marked but they are very useful for inclusion in classroom tests and for exercise purposes. However, sometimes the most straightforward completion items can cause problems in the scoring. In the following example *was preparing* and *prepared* are equally correct. It can be argued that *had prepared* is also correct if *as* is regarded as meaning *because*: i.e. people gasped because they didn't expect him to prepare for the journey – they thought he would go without preparation.

PREPARE 1. He heard a gasp behind him as he . . . . . to go.

Unexpected ways of completing blanks are shown yet again in the following example:

As soon as possible the next day I sent my story . . . . . the editor . . . . . the magazine . . . . . which my best work usually appeared.

It is quite possible to write a story *about* an editor and send the story about the editor *to* a magazine. Although such an interpretation may sound somewhat absurd, it illustrates the lengths to which the test writer must sometimes go to make certain that testees produce only the answer he or she wants to be used in each blank. For class tests, such a critical attitude might well be harmful if it took the teacher's attention off more important and urgent problems in teaching and testing.

The following example[2] indicates the wide range of possibilities for one completion item:

I go to the cinema regularly, but I . . . . . to the theatre for months.

The answer obviously required by the tester is *haven't been*; however, possible answers are:

haven't been	shan't be going
hadn't been	can't go
(sometimes) don't go	haven't been able to go
may not go	am not going
don't know whether I've been	didn't go
shan't go	haven't gone
won't go	haven't been going

If the aim of this particular item is to force the use of the present perfect tense, there are three ways of restricting the choice available to the testees (although the first two ways depend heavily on reading comprehension):

(a) *by providing a context:*

Kim usually goes to the cinema about once a week but she . . . . . four films already this month and it's only the 20th today. (Possible answers: has seen/will have seen/must have seen)

(b) *by providing data*:

I go to the cinema regularly, but it's ages since I last saw a play.
I go to the cinema regularly, but I . . . . . to the theatre for months.
(Possible answers: haven't been/haven't gone/haven't been going/haven't been able to go)

(c) *by using multiple-choice techniques:*

I . . . . . to the theatre three times since I last saw you.

A. go C. had gone
B. have been D. went

There are two major advantages in using a passage of continuous prose rather than separate sentences when giving a completion type test. Firstly, the use of context often avoids the kinds of ambiguity referred to in the previous paragraphs. Secondly, the students experience the use of grammar in context, being required to use all the context clues available in order to guess many of the missing words. As a consequence, they are generally advised to read or glance at the whole passage before starting to fill in any of the blanks.

The following are examples of two different types of completion items in context. In example (a) the blanks are indicated while in example (b) (sometimes referred to as a modified cloze passage) the omissions themselves are not indicated. In example (a) only one word should be written in each blank, while in example (b) one word has been omitted from each line. In this latter type of example, the students are required to put an oblique stroke (/) at the place where the word has been omitted and then to write the missing word in the appropriate space.

(a) It (1) ..... always useful
(2) ..... practise answering
the types of questions (3) .....
you may (4) ..... asked. However,
(5) ..... is not enough
simply (6) ..... glance through
a past paper (7) ..... answer
the questions (8) ..... your head.
This will give you (9) ..... indication
at all (10) ..... what you can do.

(1) .....
(2) .....
(3) .....
(4) .....
(5) .....
(6) .....
(7) .....
(8) .....
(9) .....
(10) .....

(b) Historians and anthropologists used think
that ill health and disease prevalent among
primitive tribes. Results recent investigations
however, have shown we much to learn
from primitive people. Without benefit of
immunisation or vaccination, primitive often
acquired immunity the diseases present in
their society. Moreover, the diseases now beset
modern society were uncommon primitive
man. Although he did live a long life,
according to research, he have been far
healthier previously thought.

(1) .....
(2) .....
(3) .....
(4) .....
(5) .....
(6) .....
(7) .....
(8) .....
(9) .....
(10) .....
(11) .....
(12) .....

In the following example, part of a sentence has been omitted. Although sentence completion items of this type make concentration on specific grammatical points more difficult, they nevertheless offer a useful way of testing an ability to handle structures according to certain patterns. Whether such items are regarded as tests of grammar or of controlled writing is of minor importance: they demand an ability to use appropriate grammatical forms for a particular purpose in a particular context.

Many zoo officials are worried by the increasing ill-treatment of animals by visitors to the zoo, especially by children. Most visitors expect to see the animals very active in their enclosures. When the animals fail to perform in the ways expected, ...
Crocodiles seem to be the chief victims of such attacks because they ..... The bottles, cans, sticks and stones that ..... end up as unsightly litter in the enclosures. Some animals, however, swallow ..... Most of the visitors who ..... are not really malicious but simply ..... Zoo officials are constantly ..... Indiscriminate feeding of the animals is not allowed because .....

The completion of dialogues can also provide a useful way of testing the ability to manipulate the grammar and patterns of the language in context. Again, meaning plays a key role in determining the students' ability to provide satisfactory answers. Unfortunately, however, it is often very difficult to write natural dialogues and at the same time provide students with useful cues.

A: .....
B: So do I. I generally watch it for an hour or two every evening.

A: .....
B: Immediately after I've finished my homework, at about eight-thirty or nine.

A: .....
B: So are mine. The funnier, the better. I like American ones best of all – you know, programmes like 'Benson' and 'Different Strokes .

A: .....
B: Neither do I. I can't stand anything that's too frightening.

A: .....
B: I agree about educational programmes, but there are still very few of them.

4.7 Constructing transformation items

The transformation type of item is extremely useful for testing ability to produce structures in the target language and helps to provide a balance when included in tests containing multiple-choice items. It is the one objective item type which comes closest to measuring some of the skills tested in composition writing, although transforming sentences is different from producing sentences. Subjective decisions, of course, may have to be made in the scoring process: e.g. how should spelling errors be marked?

The following transformation items have been based on errors which occurred in the student's letter, an extract of which was given in Section 4.3.

Rewrite each of the following sentences in another way, beginning each new sentence with the words given. Make any changes that are necessary but do not change the general meaning of the sentence.

1. I haven't written to you for a long time.
   It's a long time .....
2. In sunny weather I often go for a walk.
   When the weather .....
3. Old Mr Jones likes to look at the children playing.
   Old Mr Jones enjoys .....

Other transformation items giving some idea of the range of areas that can be covered in this way are:

1. It was impossible to work under those conditions.
   Working .....
2. I don't think it's necessary for you to stay any longer.
   I don't think you .....
3. I was able to leave the office early yesterday.
   It was possible .....
4. Joe can sing better than you.
   You cannot .....
5. This book is too big to go on any of the shelves.
   This book is so big .....
6. Frances is very good at tennis.
   Frances plays .....
7. Poor Peter was bitten by a mosquito.
   A mosquito .....

8. 'When will you visit London?' Mr Strong asked me.
   Mr Strong asked me .....

As with completion items, it is often difficult to restrict the number of possible answers. However, such restrictions are not essential for constructors of classroom tests, provided that they are fully aware of all the possible correct answers and of the specific area they are testing. The following examples indicate some of the alternatives possible for four of the preceding items:

I haven't written to you for a long time.
It's a long time *since I (last) wrote (to) you*
since you received a letter from me, etc.

I don't think it's necessary for you to stay any longer.
I don't think you *need (to) stay any longer* = expected answer
will find it necessary to stay any longer = possible answer

Joe can sing better than you.
You cannot *sing as well as Joe* = expected answer
sing better than Joe = possible answer

Frances is very good at tennis.
Frances plays *tennis very well/extremely well,* etc.
= expected answer
very good tennis = possible answer

Unfortunately, the alternatives in the last three examples defeat the purpose of the tests as they stand at present, since students can avoid using the actual grammatical patterns being tested. However, it is a simple matter to rewrite them as follows:

Is it necessary for us to stay any longer?
Need .....?

You cannot sing as well as Joe.
Joe can sing .....

Frances is very good at tennis.
Frances plays tennis .....

Sometimes it is difficult to elicit the particular form we wish to test. For example:

I feel miserable even though I shouldn't.
I know I shouldn't feel miserable but *I do.*

Although *I do* is the answer required, we could scarcely fault:

I know I shouldn't feel miserable but *I certainly don't feel happy.*
I know I shouldn't feel miserable but *I am miserable.*

In some tests,[3] students may be required to rewrite a sentence beginning with a certain word underlined in the original sentence. For example:

They believed that <u>the earth</u> was flat.
The earth was believed to be flat.

This item type is a useful variation of the previous type discussed, but sometimes restricts the kind of transformation possible since the first word

of the new sentence has to appear in the original sentence. Thus it becomes impossible to test the required transformation of a sentence like *Is it necessary for us to stay any longer?* (= *Need we stay any longer?*)

Transformation can also be effected by requiring students to substitute a given verb in a sentence,[3] the new verb necessitating a change in the structural pattern.

Ten lessons *make up* the course. (CONSIST)
The course consists of ten lessons.

I *couldn't* go swimming yesterday. (ALLOW)
I wasn't allowed to go swimming yesterday.

As with all the types of items treated in the previous sections, the transformation type of item is improved if it can be put into a context (i.e. if the sentences for transformation can be written in sequence to form part of a continuous piece of prose). Unfortunately, however, the provision of a context does not allow for the range of sentence patterns the test writer may wish to test. Moreover, most students tend to treat each sentence in isolation for the purposes of the test.

The following examples illustrate how each of the sentences for transformation can be made to form part of a continuous sentence.

1 *Changing sentences according to a given pattern*

(a) Very few objective tests allow for choice.
You have .....

(b) However, the instructions should be carefully checked.
However, you .....

(c) Different types of questions on the same paper will necessitate changes in the instructions.
The instructions .....
.....

2 *Changing sentences by using selected words*

(a) Remember that it is not necessary to answer the questions in the order set. (NEED)
.....

(b) You are advised to check your answers carefully after each question. (ADVISABLE)
.....

(c) Most teachers also recommend you to leave five minutes spare at the end of the examination in order to check your paper. (SUGGEST)
.....

4.8 Constructing items involving the changing of words

This type of item is useful for testing the student's ability to use correct tenses and verb forms. It is a traditional type of test but the layout is improved in this particular case by providing blanks on the right of the text for completion. The continuity of the text is not impaired more than necessary by having both blanks and underlined words inserted in the sentences. Thus the risk of obscuring the meaning of the text is reduced.

1 *Verbs: tenses*, etc.

Researchers (1) to convince that a drug (1) .....
they (2) to test can improve the memory and that (2) .....

it (3) to be the forerunner of other drugs which (3) .....
eventually (4) to improve mental ability (4) .....

2 *Word building*

Students who were given the drug for a fortnight did considerably (1. well) in tests than others. The tests (1) .....
included the (2. memorise) of lists of words as well (2) .....
as of (3. inform) from two messages transmitted at (3) .....
the same time. During the first week there was no
(4. notice) difference between the two groups, but (4) .....
after a fortnight the group on the drug was found to
have increased its (5. able) to learn by almost (5) .....
twenty per cent.

4.9 Constructing 'broken sentence' items

This type of item[4] tests the student's ability to write full sentences from a series of words and phrases, and thus does not allow the test writer to concentrate exclusively on testing those particular grammatical features which may have just been practised in class. It is nevertheless a useful device for testing grammar provided that the tester is aware that several other areas of the language are being tested in addition to those on which he or she wishes to focus attention. Indeed, this very fact may be considered an advantage. So many students are able to score high marks on grammar items when each item is set in isolation and concentrates on only one area of grammar. Errors are made, however, when the attention of the student is concerned with the meaning of the context as a whole and with performing a number of different grammatical tasks necessary to achieve that meaning.

When setting this item, make sure that the instructions are very clear indeed and provide one or two examples. Students unfamiliar with this particular item format frequently have difficulty in knowing exactly what to do, especially as the previous experience of many may lead them to think that the presence of an oblique stroke indicates the omission of a word instead of signalling merely a fragmented sentence or series of notes. In the rubric, students should be instructed to make whatever changes are necessary to form good sentences, adding articles, prepositions, etc. where required and putting verbs in their correct tense.

. *Take* / drugs and stimulants / *keep* awake / while *revise* examination / often *be* very harmful./ It *be* far better / *lead* / balanced life / and *get* enough sleep / every night. / There / *be* / limit / degree and span / concentration / which you *be* capable / *exert*. / Brain / *need* rest / as much body./ Indeed, / it *be* quality / than quantity work / that *be* important.

4.10 Constructing pairing and matching items

This type of item usually consists of a short conversation: e.g. a stimulus in the form of a statement or question followed by a response often in the form of a statement. It is used to test the ability to select appropriate responses to stimuli which would be presented orally in normal everyday situations. The item is more useful for testing students' sensitivity to appropriacy and their awareness of the functions of language rather than their knowledge of grammar (although grammatical clues may prove important in completing this item satisfactorily). To perform the task required, students are simply required to write the letter of the correct response in the space provided.

Column 1	Letter	Column 2
Going to see a film tonight?	F	A. No, I didn't.
How was the film?	.....	B. Most are, I think.
I can't stand war films, can you?	.....	C. It's one of the reasons.
So you went to the cinema.	.....	D. I had a lot of work to do.
Don't you find war films too violent?	.....	E. Actually, I quite like them.
Have you ever seen a Japanese war film?	.....	F. Yes, I probably will.
I like war films.	.....	G. No, I haven't.
Is everyone going to see the film?	.....	H. What a good idea! I prefer them to war films.
What about going to see a cowboy film instead?	.....	I. So do I.
Why didn't you come with us to see the film?	.....	J. All right. Nothing special.
Is that why you don't like war films?	.....	K. Not really. I quite like them.

It should be remembered, of course, that this is not authentic discourse. However, although the language and situation here are inevitably artificial, the item does serve to help students to associate the language they are learning with real-life situations, albeit to a limited extent.

4.11 Constructing combination and addition items

These objective-type items have long been used in past tests. They should be used sparingly, however, as they involve largely mechanical responses on the part of the student. Note that although the separate sentences are linked to one another by theme, the items can hardly be described as being contextualised in any real way.

1 *Combination items*
(Students are instructed to join each pair of sentences, using the word in brackets.)

(a) You finish the paper. Then check your answers carefully. (AFTER)
(b) Some questions may be very difficult. They should be left until later. (WHICH)
(c) You should usually write answers in complete sentences for all the questions on your paper. However, write notes for those questions which you do not have time to answer. (ALTHOUGH)

2 *Addition items*
(Students are instructed to insert the word in capitals in the most appropriate place in each sentence.)

(a) YET — Have you answered all the questions?
(b) STILL — Some students had not mastered the correct techniques for answering examination questions.
(c) OCCASIONALLY — There may be little choice of questions.

Notes and references

1 Similar types of test items appeared in past papers of the Joint Matriculation Board: *Test in English (Overseas)*.
2 I am indebted to Mr John Bright for this example and the possible solutions.
3 Similar items appeared in the past in the University of Cambridge Local Examinations Syndicate: *Certificate of Proficiency in English*.
4 Similar items have appeared in the University of Cambridge Local Examinations Syndicate: *First Certificate in English*.

5

Testing vocabulary

5.1 Selection of items

A careful selection, or sampling, of lexical items for inclusion in a test is generally a most exacting task. Many of the more traditional types of vocabulary tests are designed in such a way that they test a knowledge of words which, though frequently found in many English textbooks, are rarely used in ordinary speech.

The first task for the writer of a vocabulary test is to determine the degree to which he or she wishes to concentrate on testing the students' active or passive vocabulary. The next task is to decide whether the lexical items in the test should be taken from the spoken or the written language. Selection of vocabulary can thus be thought of as falling into the following rough divisions according to the four major language skills:

Listening: passive/spoken
Reading: passive/written
Speaking: active/spoken
Writing: active/written

All four divisions can be included in a single test, of course, but even then careful consideration should be given to the different weighting each division will carry in the test: for example, should there be a greater concentration on those lexical items selected from the students' reading material? Generally speaking, the more elementary the level of the test, the greater the number of lexical items associated with the spoken language.

The test constructor's task is made much easier if all the students have followed a particular syllabus. Lexical items can then be selected from:

- the syllabus (including a word frequency list if available);
- the students' textbook (provided the items approximate to those used in natural speech situations);
- the students' reading material (e.g. simplified readers, literary texts); and
- lexical errors taken from students' free-written work (or from students' incorrect answers in a cloze test).

The following error, however, may be one of verb patterning or simply the wrong choice of verb:

Is the government going to contribute the new industry?

If an error of verb patterning, the correct version would be

Is the government going to *contribute to* the new industry?

If caused by the wrong choice of verb, it would be

Is the government going to *subsidise* the new industry?

Moreover, according to the findings of research[1] conducted into the effectiveness of distractors in multiple-choice vocabulary tests, those distractors based on students' incorrect answers in cloze tests (though moderately useful) were found to be less powerful than

(a) the use of false synonyms (i.e. words with equivalent meanings to the key word underlined or shown in italics in the sentence but inappropriate in the particular context):

I'd like to book two . . . . . in the circle, please.
seats = correct word
chairs = false synonym

(b) contextually relevant items (i.e. words related to the context but different in meaning to the key word in the sentence):

'How much is a . . . . . to Tokyo, please?'
'Nine hundred yen, and a return is sixteen hundred yen.
single = correct word
ticket = contextually relevant

The test constructor is faced with a difficult problem if the testees have followed different syllabuses. Such a situation is generally associated with proficiency tests, in which a student's suitability and potential for a certain task are tested (e.g. university studies in the medium of English). In these cases, the tester may wish to base the selection of lexical items on those used in the tasks for which the student is being tested. An alternative method, appropriate for all kinds of tests, is the selection of items from such well-known word lists as *A General Service List of English Words* (Michael West – Longman), *Cambridge English Lexicon* (Roland Hindmarsh – Cambridge) and *The Wright Frequency Count*. These (and other) word lists, however, are based entirely on the written language; furthermore, no account is taken of difficulty levels (as opposed to frequency levels) and of areas where interference is encountered between the vernacular and the target language.

But testing the extent of a student's vocabulary is only one aspect of the whole problem: control of the vocabulary at his or her disposal must also be measured. An ability to discriminate between words is of the utmost importance at all but the elementary levels. In broader terms, this ability to discriminate may be regarded as developing a *feel* for the language.

Tests of vocabulary should avoid grammatical structures which the students may find difficult to comprehend. Similarly, tests of grammar should contain only those lexical items which present no difficulty to the students.

5.2 Multiple-choice items (A)

It is useful to distinguish between the following two major kinds of multiple-choice vocabulary items:

Group A Choose the letter of the word which is the nearest in meaning to the word in italics.

He's been very *feeble* since his illness.

A. unwell B. thin C. foolish D. weak

Group B Choose the letter of the correct or best word to complete each sentence.

Have you heard the planning committee's . . . . . ror solving the city's traffic problems?

A. theory B. design C. propasal D. purpose

This section concentrates on Group A items and the next section on Group B. The following item types are examples of four vocabulary recognition items which fall within the first group.

Type 1 In this type of recognition item the stem is replaced by a picture. The testees see the picture and have to select the most appropriate word relating to the picture from four or five options. This type of item is clearly very appropriate at the elementary stages.

A. running
B. jumping
C. standing
D. kicking



Type 2 Here the stem consists of a definition: the testees have to select th correct option to which the definition refers.

a person who receives and pays out money in a bank

A. broker B. accountant C. creditor D. cashier

Type 3 The stem consists of a lexical item: the testees have to select the best synonym or definition.

advocate

A. support B. advise C. contradict D. damage

dilatory

A. growing gradually larger
B. slow in getting things done
C. showing care and effort
D. heavy with drops of water

Type 4 The stem here consists of a sentence. Hence, this type of recognition item is generally to be preferred to the previous three types in so far as the 'problem' word appears in context. Vocabulary is much more usefully tested in context since it is the context that gives specific meaning and relevance to a word, thus creating a situation which is as linguistically valid as possible in the circumstances.

It's rained *continuously* for two whole days.

A. without stopping
B. heavily
C. regularly
D. at odd moments

Since subtle shades of meaning are often determined only by the specific context in which a particular word appears, it is generally advisable to provide fairly full contexts for vocabulary testing, especially at an advanced level. The fuller the context, however, the more difficult it sometimes can prove to find plausible distractors. Few good distractors, for example, can be found for the following item:

We've had to *put off* the meeting until next week. (postpone)

Synonyms are not always interchangeable in a context (without altering the meaning). However, where a word may be replaced by another in a particular context, testees may easily be misled into regarding synonyms as being generally interchangeable.

Guidelines for writing items

1 If the problem area being tested is located in the options (as in Type 2), the stem should be kept simple. If, however, the problem area is included in the stem (as in Types 3 and 4), the options themselves should be simple in so far as they should contain only those vocabulary items which the testees can understand.

2 Each option should belong to the same word class as the word in the stem, particularly when the word appears in the context of a sentence. If this rule is observed, there will be less danger of the context providing important grammatical clues for the testees. For example, although the first of the following test items is usable, options A, B and C in the second item would be grammatically incorrect when put in the context.

contemptuous

A. deep in thought
B. without a sense of humour
C. self-satisfied
D. scornful

Ian was *contemptuous* of the efforts of his friends to raise some money for the charity.

A. deep in thought
B. without a sense of humour
C. self-satisfied
D. scornful

3 The correct option and the distractors should be at approximately the same level of difficulty.[2] If the correct option is more difficult than the distractors, the testees will arrive at the correct answer by process of elimination. Thus, the test may have a negative effect on the testees: i.e. they will select the correct option not because they know it is correct but only because they know the other options are wrong. The following item measures the testees' knowledge of the distractors rather than their familiarity with the correct option:

theatrical

A. angry B. histrionic C. proud D. foolish

The converse also holds good. If the distractors are more difficult than the correct option, the item may be equally unreliable. In such a case, there will usually be a tendency for the more able students to think that the correct option is too easy and therefore wrong; they are thus tricked into selecting one of the more difficult options:

suffice

A. be adequate B. harass C. acquiesce D. be contrite

4 There is some disagreement concerning the relationship of the options t the problem area being tested. Some test writers argue that the options should be related to the same general topic or area, while others prefer as wide a range of associations as possible. Unless the vocabulary item being tested has a very low frequency count (i.e. is very rarely used), however, the item writer is advised to limit the options to the same general area of activity where possible.[2]

Item 1	*Item 2*
apparition	apparition
A. skeleton	A. scenery
B. ghost	B. ghost
C. nightmare	C. magician
D. corpse	D. castle

If item 2 were set in a test, students who had read a few ghost stories would probably select option B because they would associate *apparition* with the stories they had read. In item 1, however, students are required t show a much greater control over vocabulary.

5 All the options should be approximately the same length.[2] There is a temptation both in vocabulary and in reading comprehension tests to mak the correct option much longer than the distractors. This is particularly true in a vocabulary test item in which the options take the form of definitions: the item-writer tends to take great pains to ensure that the option is absolutely correct, qualifying it at great length. However, the item-writer rarely takes such trouble over the distractors, since they are deliberately wrong and need not be qualified in any way.

a hitch-hiker

A. a man who makes ropes
B. a person who travels about by asking motorists to give him free ride
C. an old-fashioned sailor
D. a boy who walks long distances

Any student who did not know the meaning of *hitch-hiker* would clearly choose option B – and would be correct in doing so. Consequently, if it is ever necessary to qualify a definition at some length, either one distractor or all three or four distractors should be made equally long. In this way, the correct option will be disguised a little more effectively.

It is advisable to avoid using a pair of synonyms as distractors: if the testees recognise the synonyms, they may realise immediately that neither is the correct option, since there can be only one correct answer.

The old woman was always *courteous* when anyone spoke to her.

A. polite B. glad C. kind D. pleased

Even such near synonyms as *glad* and *pleased* are sufficient to indicat to intelligent students that the choice must be between *polite* and *kind*, since if *glad* were correct, *pleased* would probably also be correct.

It is also dangerous to 'pair off' options by providing an antonym as distractor. Options A and C in the following vocabulary item immediatel stand out; again, clever students will be able to narrow their choice down to two options once they realise that A means the opposite of C.

ascend

A. go up B. talk C. come down D. fetch

5.3 Multiple-choice items (B)

The guidelines given in 5.2 for constructing vocabulary items apply equally for the Group B items now being treated. In certain ways, the items shown in this section are more difficult to construct than those in the previous section. The problem is chiefly one of context: too little context is insufficient to establish any meaningful situation, while too much context may provide too many clues (both grammatical and semantic).

1. I saw a nasty . . . . . between two cars this morning.
   A. happening B. danger C. damage D. accident

2. I was speaking to Cathy on the phone when suddenly we were . . . . . .
   A. hung up B. run out C. broken down D. cut off

3. I should have returned this book last Tuesday: it is now five days . . . . . .
   A. postponed B. excessive C. overdue D. delayed

4. Nothing had been organised and confusion seemed . . . . . .
   A. inevident B. inefficient C. ineligible D. inevitable

5. Tom always tries to help people, but recently he has been . . . . . kind and generous.
   A. chiefly B. especially C. principally D. fundamentally

Many multiple-choice vocabulary test items of the type being dealt with in this section rely on the context itself to provide grammatical clues which automatically rule out at least one of the options. These kinds of test items are useful in many respects but may possibly belong more to tests of grammar and structure rather than to vocabulary. Nevertheless, there can be little objection to introducing, say, a few items on verb patterning in a test of vocabulary.

6. I'm . . . . . of getting a new job: I don't like my present one.
   A. contemplating B. thinking C. desiring D. hoping

7. Ann . . . . . me of a girl I used to know.
   A. recalls B. reminds C. remembers D. recollects

It is sometimes argued that many multiple-choice vocabulary tests consist largely of items such as the following and that these test only a knowledge of collocation.

8. The television station was . . . . . with letters and phone calls after the announcement.
   A. drowned B. stormed C. deluged D. absorbed

Since this item ignores the ability to create unexpected collocations, it can also be argued that an imaginative use of the language is discouraged. Although there may be some truth in this argument, unexpected collocations result from a creative and intuitive handling of language, which in turn demands an implicit understanding of everyday collocations. It is usually the writer's very awareness of the degree of incongruity which makes a new collocation vigorous and meaningful.

Although the collocations in such items as the following may be tested equally well without a context, it is usually advisable to test them in sentences.

9. Dr Heston charges a high . . . . . for his services.

A. fee B. profit C. salary D. payment

(Collocations being tested here, for example, are: charge a fee/make a profit/receive a salary/make or receive a payment – although it is possible to charge a payment to an account.)

10. I don't believe you: I think you're . . . . . lies.

A. saying B. talking C. speaking D. telling

11. Iron will eventually . . . . . if grease is not applied.

A. wear B. corrode C. damage D. corrupt

12. My driving licence . . . . . at the end of this month.

A. expires B. passes out C. retires D. concludes

If separated from such contexts as the preceding ones, these test items would read:

9. charge a *fee/profit/salary/payment*
10. *say/talk/speak/tell* lies
11. iron *wears/corrodes/damages/corrupts*
12. a licence *expires/passes out/retires/concludes*

In this type of item, however, each context requires a 'normal' reaction and takes no account of cultural differences. For example, in the following item B or D would be correct in certain societies since it is impolite to accept a gift without first vehemently refusing it.

Emma cried out with . . . . . at the beautiful present Mrs White gave her.

A. delight B. horror C. dismay D. anger

In view of such ambiguity, it is even more important than usual to provide a context for this particular kind of item. The following is a typical example of this type of multiple-choice item as it appears in several tests. The dialogue takes place in a doctor's surgery which has a pharmacy.

PAUL LEE: Can you tell me what time the doctor's (1) . . . . . opens?
MRS KING: It's open now. The (2) . . . . . will help you.
PAUL LEE: Excuse me. I just want to collect a (3) . . . . . .
MRS KING: Is it for some (4) . . . . . for a headache?
PAUL LEE: No, it's for some cough (5) . . . . . .
MRS KING: Here it is. This should soon (6) . . . . . your bad cough.

(1) A. office B. surgery C. hospital D. ward
(2) A. porter B. hostess C. waitress D. receptionist
(3) A. prescription B. recipe C. cure D. direction
(4) A. prevention B. liquid C. medicine D. solution
(5) A. mixture B. drink C. wash D. compound
(6) A. prevent B. treat C. refresh D cure

This is undoubtedly the most common type of multiple-choice vocabulary item. The provision of a context, however, limits the test constructor to testing only the vocabulary associated with a particular topic. Hence, many well-known tests still tend to include single sentence vocabulary items rather than fully contextualised ones in order to cover the range of

vocabulary considered desirable for sampling. The choice between the use of single sentences and the use of paragraphs providing a far fuller context will be determined by the purpose of the test and the test writer's own approach to the communicative aspects of language learning. The question is simply whether the testing of language in context is worth the sacrifices demanded, and the answer must differ according to each particular situation.

5.4 Sets (associated words)

Many of the difficulties arising from the testing of collocations are avoided by the testing of word sets. In such tests the students' familiarity with a range of associations is measured.

Type 1: Recognition

Read each of the following lists of four words. One word does not belong in each list. Put a circle round the odd word in each list.

son	happy	arrive
father	married	depart
boy	engaged	go away
brother	single	leave

Type 2: Production

Each group of words is related to a particular subject. Write down the particular subject which is connected with each group of words.

hand	theatre	volume	nursery
wrist	sister	track	lift
dial	bed	head	slope
face	ward	spool	snow
(= watch)	(= hospital)	(= tape recorder)	(= skiing)

5.5 Matching items

Type 1 of the following test items suffers from testing lexical items from different word classes, while Type 2 tests a mixed bag of tense forms, etc. The result is that for both types of test items grammatical clues assume great importance, since they are instrumental in limiting the range of choices facing the testees for each blank. For example, although there may appear to be 20 words for selection for blank (1) in Type 1, in practice there are only three which would fit grammatically: *turned* (*down*), *broken* (*down*), *knocked* (*down*). Similarly, in the first sentence of Type 2 there are only two options (*pull through*, *get away*), since all the other options are either past tense forms or participles. Both items need to be rewritten, therefore, if a higher degree of reliability is to be obtained.

Type 1

Write the correct word from the following list at the side of each number on your answer sheet. Use each word once only.

road	accident	travelling	turned	side
broken	know	knocked	middle	looked
lorry	policeman	pavement	running	hurt
lying	crossed	left	forgot	talk

Poor Tom Wright was (1) down by a (2) last week when he was crossing the (3). He was quite badly (4) and he had to go into hospital for a few days. His left leg was (5) and both his arms were cut. While he was (6) in bed in the hospital, a (7) came to (8) to him.

'Was the lorry (9) very quickly?' he asked Tom.
Tom told him all about the (10).
'I was (11) home from school and I (12) the road. I (13) right but I (14) to look (15). In the (16) of the road I suddenly saw a lorry. I didn't (17) what to do, so I began to run to the other (18) of the road. The lorry (19) but it hit me when I was near the (20).'

Type 2

Complete the following sentences with the most suitable verb phrase from the list.

came about	pull through	broken out	falling out
running into	brought up	get away	put off

1. 'Did the prisoner manage to . . . . . ?' 'Yes, the police are still looking for him.'
2. The doctor thought Mr Benson would . . . . . after the operation.
3. The couple are always . . . . . and causing a disturbance.
4. And so it . . . . . that we eventually parted.

(etc.)

It is much more efficient to test words from the same word class (e.g. nouns only in Type 1), or parallel tense forms (e.g. the past simple tense in Type 2). Thus the Type 2 item could be rewritten as follows:

came about ran into pulled through got away

1. 'I hear the prisoner . . . . . yesterday and the police are still looking for him.'
2. 'We were all relieved that Mr Benson . . . . . after the operation.'

(etc.)

Type 3

From the list of words given, choose the one which is most suitable for each blank. Write only the letter of the correct word after each number on your answer sheet. (Use each word once only.)

A. completely	C. busily	E. quickly
B. politely	D. carefully	F. angrily

'Write (1) . . . . . 'the teacher shouted (2) . . . . . , 'but don't waste time. You must get used to working (3) . . . . . '
'Please, sir,' a student said (4) . . . . . , 'I've finished.'
'No, you haven't,' answered the teacher. 'You haven't (5) . . . . . finished until you've ruled a line at the end.' Meanwhile, the boy sitting next to him was (6) . . . . . engaged in drawing a map.

This type is satisfactory in many ways because all the lexical items tested are adverbs. However, like the other two types, this type gives the student too little choice. For instance, there will be only one word left for the last number. Thus, it could be improved considerably by the addition of a few other adverbs. The list might then read as follows:

A. completely	E. deliberately	I. quickly
B. heavily	F. busily	J. hardly
C. ably	G. hastily	K. angrily
D. politely	H. carefully	L. suitably

The first attempt to construct this list included the adverbs *silently* and *already*, but it was then found that either of these could be used at (6) instead of the correct option *busily*. This illustrates one of the dangers of this particular testing device: clearly the more distractors there are, the greater is the chance that one of the distractors might be a correct option for at least one of the other items.

Type 4

The most useful type of matching item is undoubtedly that based on a reading comprehension passage. The students are given a list of words at the end of the passage and required to find words of similar meaning in the passage. Since a detailed context is provided by the passage and little additional material is required, this is an economical method of testing vocabulary. The chief risk here, however, is the duplication of questions: if one of the reading comprehension questions depends for its answer on a knowledge of the meaning of a particular word, care must be exercised *not* to test that word again in the vocabulary section.

In most well-known tests in which this type of item is included, a different reading text from the comprehension text is used as the basis for the matching vocabulary test. Thus, the reading text contains only questions on vocabulary and does not include comprehension questions as such. In this way the test constructor can be sure that the ability to answer the comprehension questions does not depend on a knowledge of the individual words selected for the vocabulary test.

In the following example[3] candidates in the test are instructed to replace the words listed below with the appropriate words contained in the passage without changing the meaning.

groups	.....
owned	.....
specific	.....
made up	.....
chief	.....
knowledge	.....
similarly	.....
close to each other	.....
were inclined	.....
work together	.....

THE TEHUELCHES

The Tehuelches lived in a band – usually of between fifty and a hundred people. Each band had exclusive rights to a particular hunting area and no other band was able to hunt there without permission. Each band was composed of families related through the male line and the man who led them was the hunter who had the greatest experience of the hunting groups. Each man married a woman from another band and his sister would also marry men outside his band. In this way bands in a neighbourhood were linked by ties of marriage and so tended to co-operate with each other in hunting and other tasks.

5.6 More objective items

This section contains examples of types of vocabulary items which have appeared in certain tests. While Types 1 and 2 are useful for classroom testing, Types 3 and 4 are rather artificial, and should be avoided where possible.

Type 1: Word formation test items

(a) Write a word in each blank. The word you write must be the correct form of the word on the left.

(i) CARE Be . . . . . when you cross the road.
(ii) CRUEL To mistreat animals is a form of . . . . . .
(iii) INTEREST Do you think this book is . . . . . ?
(iv) ENTER Can you show me the . . . . . to the cave?

(b) Rewrite the following paragraph putting in each blank the correct form of the word in capital letters.

MOMENT

Can you spare a . . . . . ?' Peter asked his brother. He thought he could detect a . . . . . look of impatience on his elder brother's face, but it was gone in an instant.
'I'm very busy at the . . . . . ,' his elder brother said.
What is it you want to speak to me about?' he asked Peter.
Peter's mind . . . . . went blank. 'I've forgotten he said.
'Well, then it must have been nothing of . . . . . importance,' his elder brother said rather sarcastically.

Type 2: Items involving synonyms

(a) Write in each space the best word to replace the words underlined in each sentence.

(i) Tom went <u>at once</u> to the doctor's. immediately
(ii) <u>All of a sudden</u> there was a loud cry. . . . . .
(iii) I <u>came across</u> an interesting book. . . . . .
(iv) The boat is over fourteen feet <u>in length</u>. . . . . .

(b) In each space write one word that means almost the same as the word on the left. The word you write must rhyme with the word on the right.

Example: early *soon* moon

(i) purchase . . . . . die
(ii) miserable . . . . . bad

A similar item may be constructed so as to involve antonyms rather than synonyms. The phonological element (rhyming) in 2(b), however, may only confuse testees instead of helping them. Words are tested in isolation, so, apart from its sheer novelty, the item is of little use and is not to be recommended for most purposes. The activity involved is more a game than a test

Type 3: Rearrangement items

Rearrange the following letters to make words. Then use each word in a sentence of your own so as to show the meaning of the word.

PLEAP ROLRY CELPA
SUHOE IRACH EGURA

As can be seen, this item is little more than a crossword puzzle. It may, perhaps, be of some use in an intelligence test, but it is of doubtful use in a language test.

Type 4: Definitions

(a) Use each of the following words in a sentence so as to show the meaning of the word.

economy politics industrious (etc.)

(b) Explain the meaning of each of the underlined words in the following phrases.

an archaic word a fortuitous event

These item types are of very little use. They test writing ability in addition to a knowledge of word meanings. Furthermore, it is extremely difficult even for native speakers to produce sentences 'to show the meaning' of words – and it is certainly not a useful task. A student may be familiar with the meaning of a word and may use it correctly, without being able to express this meaning clearly in a sentence (especially under test conditions).

5.7 Completion items

The following types of completion items can be used for the testing of vocabulary. Tests which present such items in a context are generally preferable to those which rely on single words or on definitions.

Type 1

Read through the following passage containing a number of incomplete words. Write each completed word on your answer sheet at the side of the appropriate number. (Each dash represents one letter.)

Snakes are one of the (1) d–m–n––t groups of (2) r–pt––––: there are at least 2,000 different (3) sp–c––s of snakes (4) sc–t––––d over a wide area of the earth. Not all snakes are (5) p––s–n––s: in fact, the (6) m–j––––y are quite harmless. Contrary to (7) p–p–l–– belief, a snake's (8) f––k–d tongue is not (9) d–ng––––– to human beings: it is merely for touching and smelling (10) s–bs––n––s. Snakes (11) in––ct poison into their (12) vi––––'s body by (13) b–t––g him with their (14) f––gs.

Type 2

(a) Complete each blank with the most appropriate word to replace each number in the text.

ROSNAH: What's the (1) today? (1) .....
MOHAMED: It's the seventh.
ROSNAH: At what (2) does the concert start? (2) .....
MOHAMED: Seven o'clock, I think. Just a moment. I made a note of it in my (3). (3) .....
ROSNAH: How long do you think it'll (4)? (4) .....
MOHAMED: It finishes about ten.
ROSNAH: That's quite a long (5), isn't it? (5) .....
MOHAMED: I suppose so. It's three hours.

Note the range of possible answers, especially with 3, 4 and 5 (e.g. 3: diary, notebook, exercise book; 4: last, take; 5: time, concert, performance).

(b) Complete the following paragraph on problems caused by weightlessness by writing ONE word in place of each blank.

Increasing ..... is now being ..... on the effects of weightlessness on man. For ....., scientists are ..... the role of gravity on the way cells function. Even in the first manned space-flights doctors were largely unaware of the various problems .....by absence of gravity. They found that weightlessness ..... in the redistribution of blood and other fluids from the legs to the top of the body. They were able to ..... how astronauts' legs actually shrank and their faces swelled during the first few days of space flight. Moreover, doctors had a chance to ..... astronauts both at the time of the flight by ..... of television cameras and after the flight during extensive medical ..... Scientists are now ..... into the effects of diet and exercise as a ..... of reducing some of the problems ..... by weightlessness. ....., most of the fundamental scientific questions will never be satisfactorily ..... by scientists working on the Earth.

These test items come close to the kind of item often used to test reading comprehension (described in Chapter 8). Clearly, a degree of comprehension is necessary before each of the blanks can be completed. The items have been included in this chapter because there is a deliberate attempt to concentrate on the testing of vocabulary – in the first case the vocabulary associated with information about the time and date, and in the second case with the language of research and inquiry.

Notes and references

1 Goodrich, H C 1977 Distractor Efficiency in Foreign Language Testing. *TESOL Quarterly* **11**
Cohen, A D 1980 *Testing Language Ability in the Classroom*. Newbury House

2 See Harris, D P 1969 *Testing English as a Second Language*. McGraw-Hill, pp. 54–57.

3 North Western Regional Advisory Council for Further Education, April 1983, *English as a Second Language* Paper 3 C4

6

Listening comprehension tests

6.1 General

An effective way of developing the listening skill is through the provision of carefully selected practice material. Such material is in many ways similar to that used for *testing* listening comprehension. Although the auditory skills are closely linked to the oral skills in normal speech situations, it may sometimes be useful to separate the two skills for teaching and testing, since it is possible to develop listening ability much beyond the range of speaking and writing ability if the practice material is not dependent on spoken responses and written exercises.

An awareness of the ways in which the spoken language differs from the written language is of crucial importance in the testing of the listening skills. For instance, the spoken language is much more complex than the written language in certain ways, as a result of the large element of 'redundancy' that it contains. An example can be seen in the spoken question 'Have you got tc go now?', the question being signalled by the rise in pitch on *go now* and by the inversion of the word order (i.e. by both phonological and grammatical features). Thus, if the listener did not hear the question signal *Have you*, the rise in pitch would indicate that a question was being asked. If the speaker slurred over *got to*, the question would still be intelligible. In addition, meaning might also be conveyed, emphasised and 'repeated' by means of gestures, eye movements, and slight changes in breathing. Such features of redundancy as those described make it possible for mutilated messages to be understood, even though the full message is only partially heard. Furthermore, the human brain has a limited capacity for the reception of information and, were there no such features built into the language, it would often be impossible to absorb information at the speed at which it is conveyed through ordinary speech. Such conversational features as repetition, hesitation and grammatical re-patterning are all examples of this type of redundancy, so essential for the understanding of spoken messages.

What is the significance of these features for testing purposes? Firstly, the ability to distinguish between phonemes, however important, does not in itself imply an ability to understand verbal messages. Moreover, occasional confusion over selected pairs of phonemes does not matter too greatly because in real-life situations listeners are able to use contextual clues to interpret what they hear. Although listeners rely on all the phonological clues present, they can often afford to miss some of them.

Secondly, impromptu speech is usually easier to understand than carefully prepared (written) material when the latter is read aloud. Written tests generally omit many of the features of redundancy and impart information at a much higher rate than normal speech does. Consequently, it is essential to make provision for restating important points, rewriting and rephrasing them when writing material for aural tests. The length of the segments in each breath group should be limited during delivery, for the longer the segment the greater the amount of information and the greater the strain on the auditory memory. The pauses at the end of each segment should also be lengthened to compensate for the lack of redundant features.

Although not always possible when listening comprehension tests are conducted on a wide scale, it is helpful if the speaker can be seen by the listeners. However excellent the quality of a tape recorder, a disembodied voice is much more difficult for the foreign learner to follow. In practice, most tape recorders are not of a high quality and are used in rooms where the acoustics are unsatisfactory. If the quality of the reproduction is poor, the test will be unreliable, especially when such discrete features as phoneme discrimination, stress and intonation are being tested.

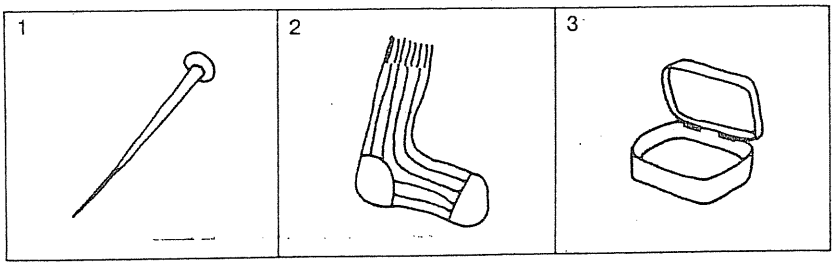
Apart from the use of videotape, however, the tape recorder is the only way of ensuring complete uniformity of presentation and thus a high degree of reliability. It is also possible to use recordings made by native speakers and thus present perfect models of the spoken language – an important advantage in countries where native speakers are not available to administer the test. Moreover, tape recorders are essential for the production and use of authentic material.

For purposes of convenience, auditory tests are divided here into two broad categories: (i) tests of phoneme discrimination and of sensitivity to stress and intonation, and (ii) tests of listening comprehension.

6.2 Phoneme discrimination tests

Type 1

(a) This type of discrimination test consists of a picture, accompanied by three or four words spoken by the examiner in person or on tape.



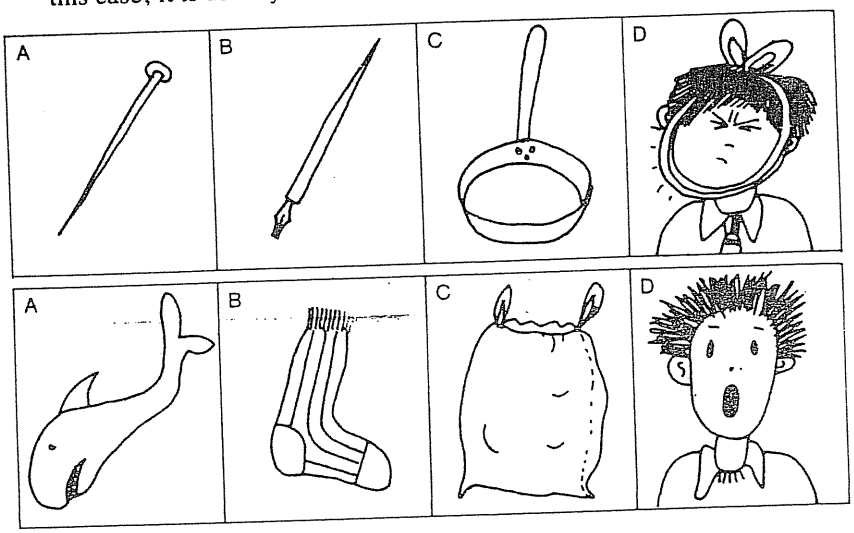
The testees hear:

1.	A. pin	B. pen	C. pair	D. pain
2.	A. shark	B. sock	C. sack	D. shock
3.	A. thin	B. tin	C. fin	D. din

After each group of four words the testees write the letter of the most appropriate word for that picture. For example:

1. A 2. B 3. B

(b) Conversely, four pictures may be shown and only one word spoken. In this case, it is usually better if the word is spoken twice.



The testees hear:

1. pain – pain (= D)
2. sock – sock (= B)

(etc.)

Type 2

The testees hear three sentences and have to indicate which sentences are the same and which are different.

1. A. There's a bend in the middle of the road.
   B. There's a bend in the middle of the road.
   C. There's a band in the middle of the road.
2. A. Is that sheet over there clean?
   B. Is that seat over there clean?
   C. Is that seat over there clean?
3. A. I've just locked the car in the garage.
   B. I've just knocked the car in the garage.
   C. I've just locked the car in the garage.

(etc.)

Type 3

(a) In each of these items one word is given on tape while three or four words are printed in the answer booklet. The testees are required to choose the written word which corresponds to the spoken word.

1. *Spoken*: den
   Written: A. ten B. den C. Ben D. pen
2. *Spoken*: win
   Written: A. when B. one C. wane D. win
3. *Spoken*: plays
   Written: A. plays B. prays C. pays D. brays

(etc.)

(b) This type of item is similar to the previous one; this time, however, the words spoken by the tester occur in sentences. The four options may then be either written or spoken.

1. *Spoken*: I'll thread it for you.
   Written or spoken: A. thread B. tread C. threat D. dread
2. *Spoken*: Did John manage to catch the train?
   Written or spoken: A. drain B. chain C. plane D. train
3. *Spoken*: Put the pan in some hot water.
   Written or spoken: A. pan B. pen C. pin D. pain

(c) This item type[1] is similiar to Type 3(a): one word is spoken by the tester (preferably twice). However, instead of a choice of four words, testees have in front of them a choice of four definitions. They have thus to select the correct definition for the word they hear.

1. *Spoken*: cot – cot
   Written: A. stopped and held
   B. a baby's bed
   C. pulled by horses
   D. a small pet animal covered with fur
2. *Spoken*: threw – threw
   Written: A. made something move through the air
   B. not false
   C. some but not many
   D. made a picture or diagram on paper
3. *Spoken*: bud – bud
   Written: A. part of a tree or a flower
   B. a creature with wings
   C. something to sleep on
   D. not good

The test items described in this section are all of limited use for diagnostic testing purposes, enabling the teacher to concentrate later on specific pronunciation difficulties. The items are perhaps more useful when testees have the same first language background and when a contrastive analysis of the mother tongue and the target language can be used. Most of the item types described are short, enabling the tester to cover a wide range of sounds.

Type 3(c), however, tests not only the ability to discriminate between the different sounds of a language but also a knowledge of vocabulary. A testee who may be able to discriminate accurately will nevertheless find the test very difficult if he or she cannot understand the definitions in the options. Similarly, Type 3(a) is a test of phoneme discrimination *and* spelling ability. In Type 3(b) proficiency in grammatical structure will favour the testee. Thus, for example, a testee who cannot discriminate between *thread*, *tread*, *threat* and *dread* may immediately rule out the distractors *threat* and *dread* since they cannot be put in the pattern *I'll . . . . . it for you.*

Each individual test item in all the types described must be kept fairly simple. Obscure lexical items should be avoided. This may seem to be a simple enough principle to observe, but the avoidance of difficult lexical items frequently makes it impossible to test all the sound contrasts that need to be included in the test. For example, the contrasts *shark*, *sock*,

sack, shock would not be suitable for inclusion in a test intended for elementary learners of English.

Much of the material in such tests is unfortunately very artificial, differing greatly from spontaneous speech. Frequently there is a tendency for the tester to adopt a certain tone-pattern and rhythm which may be a source of irritation to the listeners or affect their concentration. However, if the tester changes pitch (e.g. live, leave, live) this will only confuse the listeners. Thus, the tester must attempt to pronounce every syllable using the same stress and pitch patterns.

The ability to discriminate between certain phonemes may sometimes prove very difficult for native speakers. Many English dialects fail to make some of the vowel and consonant contrasts and thus, in addition to all the other variables (e.g. the acoustics of the room, the quality of the tape recorder, etc.), these tests are affected by the pronunciation differences of native speakers.[2]

6.3 Tests of stress and intonation

Although features of stress, intonation, rhythm and juncture are generally considered more important in oral communication skills than the ability to discriminate between phonemes, tests of stress and intonation are on the whole less satisfactory than the phoneme discrimination tests treated in the previous section. Most tests are impure in so far as they test other skills at the same time; many are also very artificial, testing the rarer (but more 'testable') features.

Type 1 The following item type[3] is designed to test the ability to recognise word stress or sentence stress. The testees listen to a sentence (usually spoken on tape) and are required to indicate the syllable which carries the main stress of the whole structure. They show the main stress by putting a cross in the brackets under the appropriate syllable.

Spoken: I've just given THREE books to Bill.
Written: I've just given three books to Bill.
( ) ( )( )( ) (X) ( ) ( )( )

Spoken: My FAther will help you do it.
Written: My father will help you do it.
( ) (X)( ) ( )( ) ( ) ( )( )

Unfortunately, this test lacks context and is very artificial. It tests only recognition of stress and is of limited use for ear-training purposes.

Type 2 The examiner makes an utterance and the testees have to select the appropriate description to indicate whether they have understood the original utterance. The utterance is spoken once only, but the test is based on the principle that the same utterance may be spoken in several different tone-patterns indicating a plain statement, a question, sarcasm, surprise, annoyance, etc.

Spoken: Tom's a fine goalkeeper.
Written: Tom's a fine goalkeeper.

The speaker is
A. making a straightforward statement
B. being very sarcastic
C. asking a question

Spoken: You will send me a couple of tickets.
Written: You will send me a couple of tickets.

This is probably
A. a request
B. a command
C. an expression of disbelief

Spoken: I'll help Ann.
Written: I'll help Ann.

The speaker is
A. reluctant to help Ann
B. eager to help Ann
C. making a plain statement

This type of test item is sometimes difficult to construct. Since the context must be neutral, it is often hard to avoid ambiguity. There is also a danger of inventing odd interpretations or of concentrating on the rarer meanings: e.g. sarcasm, irony, incredulity. Moreover, it can be argued that the item tests vocabulary and reading comprehension in addition to sensitivity to stress and intonation.

6.4 Statements and dialogues

These items are designed to measure how well students can understand short samples of speech and deal with a variety of signals on the lexical and grammatical levels of phonology. They are very suitable for use in tests administered in the language laboratory but they do not resemble natural discourse. The spontaneity, redundancy, hesitations, false starts and ungrammatical forms, all of which constitute such an important part of real-life speech, are generally absent from these types of items simply because they have been prepared primarily as written language to be read aloud.

Moreover, the responses required on the part of the listeners are not communicative responses in any sense at all. The listeners are not required to respond by interpreting what they have heard or by adding further information, as in real life. Such communicative responses, although ideal for many teaching situations, would be difficult to incorporate in such listening tests, especially those intended for particular diagnostic purposes. Nevertheless, the importance of such responses in tests of listening should be borne in mind when communicative proficiency tests are being constructed – in other words, when the test writer is interested in finding out what students can *do* with the language they are learning.

Type 1 This item type may be included in a test of grammar, a test of reading comprehension or a test of listening comprehension, depending on whether the item is written or spoken. It tests the ability to understand both the grammatical and lexical features of a short utterance. The testees hear a statement (usually on tape) and then choose the best option from four written paraphrases.

Spoken: I wish you'd done it when I told you.
Written: A. I told you and you did it then.
B. I didn't tell you but you did it then.
C. I told you but you didn't do it then.
D. I didn't tell you and you didn't do it then.

Spoken: It took Alan a long time to find he couldn't mend my bicycle.

Written: A. After a long time, Alan realised he was unable to mend my bicycle.
B. Alan spent a long time mending my bicycle but he was at last successful.
C. Alan was a long time before he found my bicycle.
D. In spite of searching for a long time, Alan couldn't find my bicycle and, therefore, couldn't mend it.

When constructing these items, it is advisable to keep the grammatical, lexical and phonological difficulties in the stem, leaving the written options free of such problems and at a lower level of grammatical and lexical difficulty than the spoken stimulus.

Type 2 These item types are more satisfactory than Type 1 insofar as they are an attempt to simulate speech situations. The testees listen to a short question and have to select the correct response from a choice of four printed ones.

Spoken: Why are you going home?

Written: A. At six o'clock.
B. Yes, I am.
C. To help my mother.
D. By bus.

Each option should be so constructed as to appear correct in some way to the testee who has not recognised the correct signals in the question. Thus, in the previous item, option A would appear correct if the testee had confused *Why* with *When*, and option D if he or she had 'heard' *How* signal the question. If, on the other hand, a testee had failed completely to pick up the Wh-question signal, he or she would be tempted to select option B, considering it the answer to a Yes/No question.

The question types should be varied as much as possible and Yes/No questions included as well as Wh-questions.

Spoken: Does Alison mind you playing the piano?

Written: A. Yes, she's always thinking about it.
B. No, she rather likes it.
C. No, she doesn't play the piano.
D. Yes, she must be careful.

In this item two of the distractors (A and D) are based on confusion relating to *mind* in order to tempt any testee who has failed to understand the question accurately. Distractor C has been included to attract any testee who has generally misunderstood the question and thinks it is about Alison playing the piano.

It can be argued that for certain tests the primary purpose of a listening comprehension item should be to test comprehension alone and not the ability to select an appropriate reply to a stimulus. It is possible that a student who fails to answer this type of item correctly may have actually understood the statement or question but failed to select the correct reply.

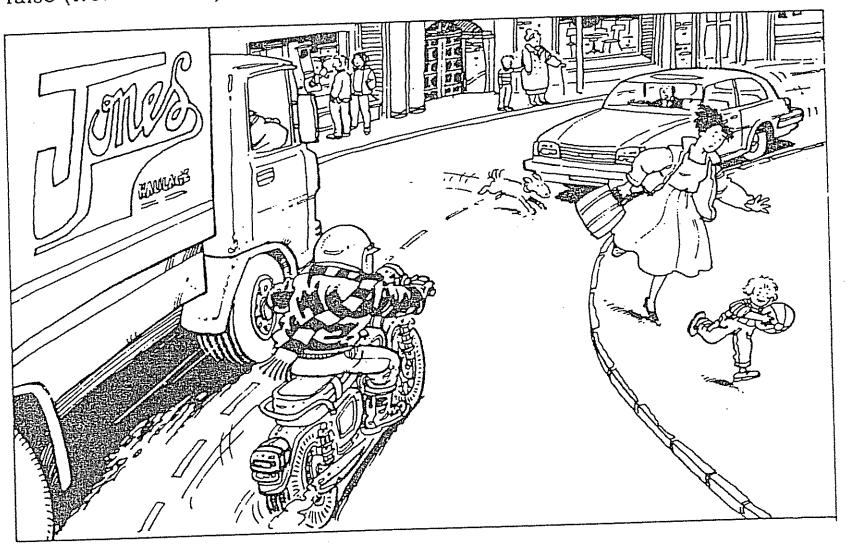
Furthermore, each statement or question which provides the stimulus in this type of item usually takes the form of an isolated item presented out of context and occurring with other unrelated items. Consequently,

students who answer all these items in a test have to accomplish an intricate sequence of mental gyrations, suddenly jumping from one situation to another. Nevertheless, this item type is useful for several purposes in class progress tests provided that the limitations indicated here are recognised and the item type is not over-used.

6.5 Testing comprehension through visual materials

Most of the item types in this section are more appropriate for the elementary stages of learning English. They are, however, preferable to the discrimination items previously discussed as they involve the testing of grammar and lexis through phonology. Pictures, maps and diagrams can be used effectively for testing such skills, thereby making the testee's performance less dependent on other skills (e.g. speaking, vocabulary and reading).

Type 1 In this item type a picture is used in conjunction with spoken statements. The statements are about the picture but some are correct and others incorrect. The testees have to pick out the true (i.e. correct) statements and write T (or put a tick ✓) at the side of the appropriate numbers. They write F (or put a cross ✗) at the side of the numbers of the false (i.e. incorrect) statements.



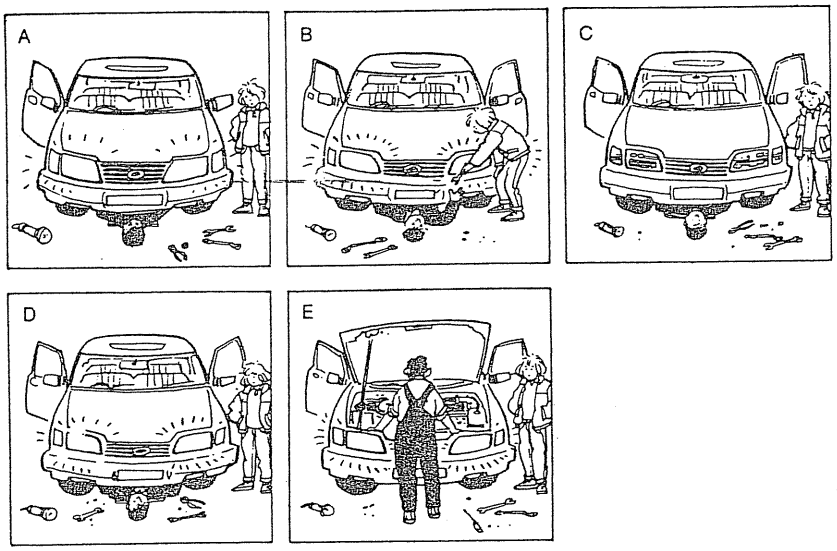
Spoken:

1. The lorry's on the left of the motorcyclist.
2. The car's travelling in the same direction.
3. A dog's running in front of the car.
4. A little girl's running after her mother.
5. She's holding a doll.
6. Her mother's carrying a bag.
7. The two boys are looking in a shop window.
8. A very small boy's helping the old woman.
9. The old woman's going into a shop.
10. A tall man's posting some letters.
11. There are a lot of cars in the street.
12. The two boys are on the same side of the street as the little girl.

Written:

1.	2.	3.	4.	5.	6.
7.	8.	9.	10.	11.	12.

Type 2 In the following listening tests students have five pictures in front of them. They listen to four sentences, at the end of which they are required to select the appropriate picture being described.



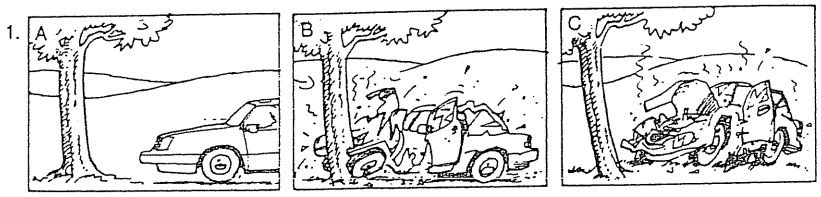
The testees hear:

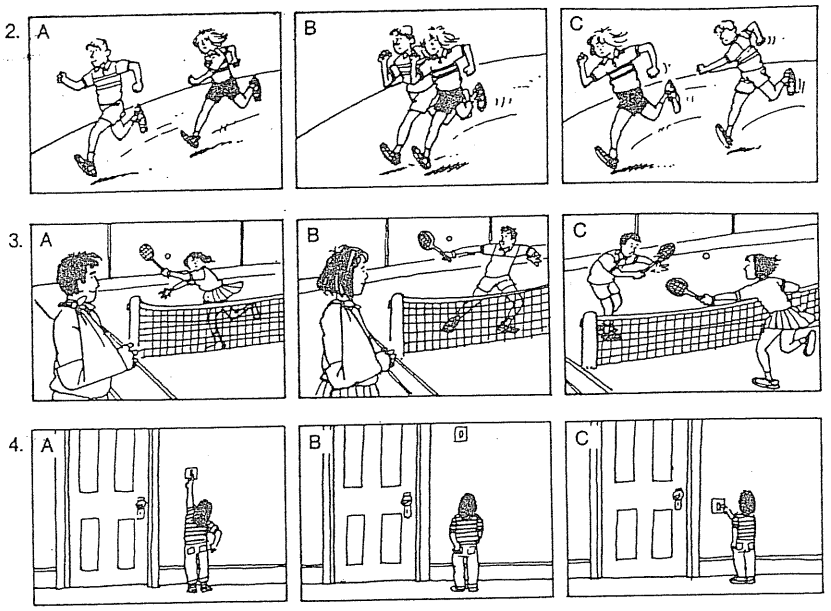
1. Both car doors are open.
2. It's daylight but both headlamps are on.
3. The man who's repairing the car is lying underneath it.
4. Although the girl sees the man working hard, she doesn't help him.

Thus the testees are able to narrow down the choice available to them as follows:

1.	B C D E	(Only A shows one door open)
2.	B D E	(Only C shows the headlamps off)
3.	B D	(Only E shows the man standing up)
4.	D	(Only B shows the girl helping the man)

Type 3 The following type of test item[4] is used in a number of listening comprehension tests. The testees see a set of three or four pictures and hear a statement (or a short series of statements), on the basis of which they have to select the most appropriate picture. In the test the testees often see a total of ten or twelve such sets of pictures.



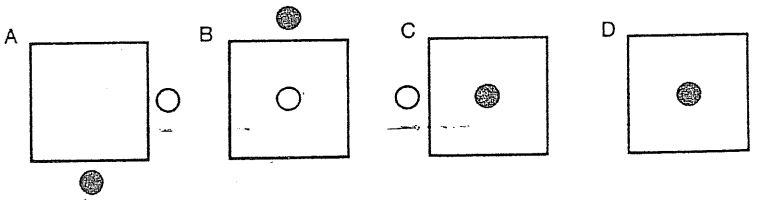


The testees hear:

1. The car's going to crash into a tree.
2. Danny can't run as fast as Claire.
3. Tom wishes his sister could play tennis with him.
4. The switch is so high that Katie can't reach it.

Type 4 Simple diagrams (consisting of lines, squares, rectangles, circles and triangles) can be drawn to function as options in a test of elementary comprehension.[5] Illustrations of this nature lend themselves in particular to testing such grammatical features as comparison, prepositions and determiners.

Look carefully at each of the four diagrams. You will hear a series of statements about each of the diagrams. Write down the appropriate letter for each statement.

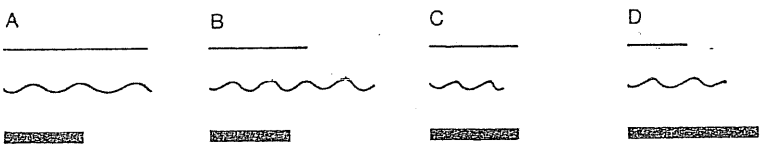


Spoken:

1. A: Look! What's that inside the square?
   B: It's a white circle.
2. A: Is that a black circle?
   B: Whereabouts?
   A: Above the square.
   B: Yes, it is. It's a black circle above the square.

3. A: Is the white circle on the left of the square?
   B: No, it's on the right of the square.
4. A: Is there anything at all in the square?
   B: No, it's completely empty. There's neither a white nor a black circle in the square.
5. A: There's nothing at all under the square, is there?
   B: No, you're wrong. There's a black circle under the square.
6. A: What are you looking at?
   B: I'm looking at a square.
   C: Which square?
   D: The one under the black circle, of course.
7. A: Is the circle on the left of the square?
   B: No, it isn't. The square's on the left of the circle.
8. A: What's unusual about this drawing?
   B: Well, there's a black circle inside the square.
   A: That's not unusual. There are two squares with black circles inside.
   B: But this one I'm looking at has a white circle just outside the square.

All kinds of shapes and forms can be used to test listening comprehension. The following example illustrates how an understanding of complex structures can be tested in this way. However, there is often a temptation for the test writer to be too 'clever' and set an item testing intelligence (or mental agility) rather than language acquisition – as in the following example.

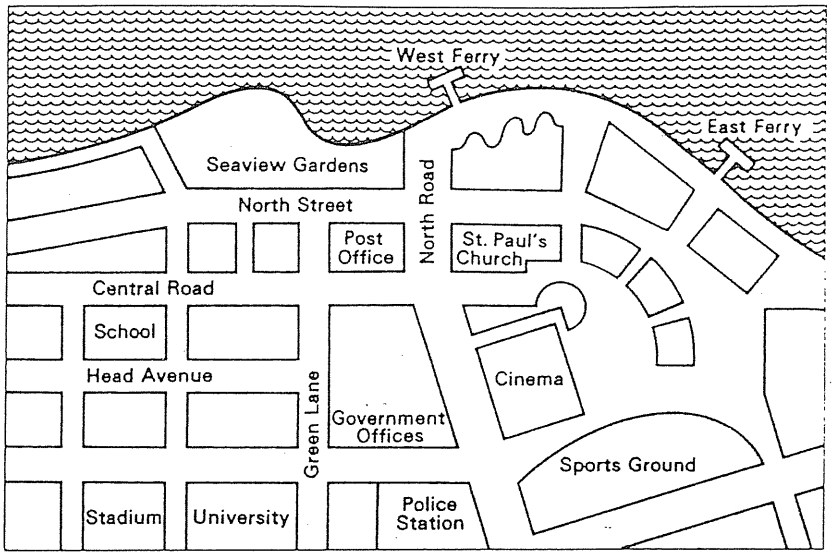


Spoken: If the thick line had been only a millimetre longer, it would have been the longest of the three lines.

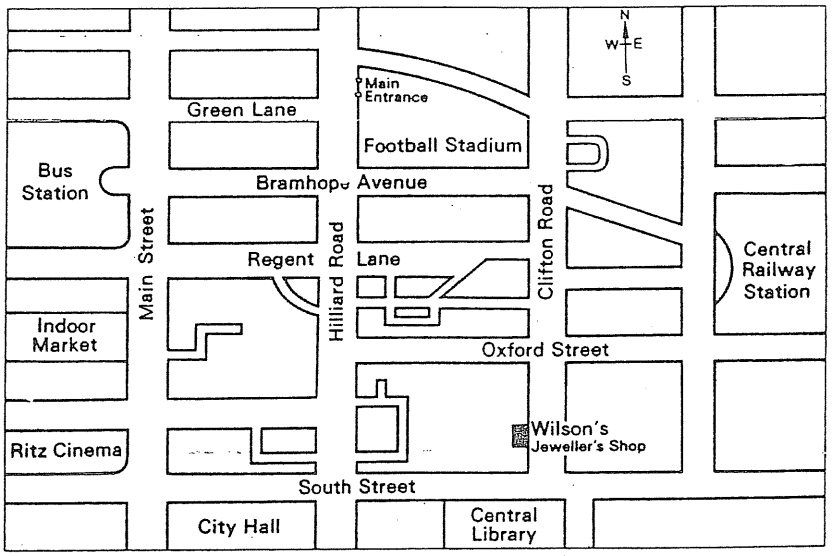
Short conversational exchanges as in the first example of Type 4 are far preferable to single sentences (as in Types 1, 2 and 3). All four types, however, are scarcely valid tests of the ability to understand natural discourse. Nevertheless, such items are of some use for certain purposes and are usually quite reliable guides to particular aspects of the listening ability.

Type 5 This test is designed to assess the testee's ability to understand simple instructions. Any street map can be used or adapted for this purpose (see the example at the top of the next page).

Spoken: You come out of school into Central Road and walk in the direction of Green Lane. However, you take the left turning just before you reach Green Lane. At the end of the street you turn right and continue until you come to the second turning right. You cross this road and you will see on your right . . . . . (Which building will you see?)



The following listening comprehension test is in the form of a dialogue. The idea on which it is based was suggested by an actual robbery and a police chase. In this way, the listener is given a greater sense of realism and an added interest in the dialogue.



A: Have you heard about the raid on the jewellery shop in Clifton Road?
B: Yes, in fact, I saw part of the chase. It was extraordinary.
A: I've only heard a very vague report about it. What exactly happened?
B: Well, the thieves planned to rob the shop – you know, Wilson's in Clifton Road – just after it'd opened early yesterday morning.

A: In broad daylight? I didn't know that.

B: They planned to arrive as the jewellery was being taken from the safe into the big display window. They arrived in a large red car which they parked on the opposite side of the road. Can you see the place on this map I've got?

1. X: Write the letter A on your map at the place where the thieves parked the car.

A: How many robbers were there?

B: Three. One waited inside the car and the other two walked over to the shop, carrying large briefcases. Once they were in the shop, they made the manager and his assistant lie down on the floor while they filled the briefcases with jewellery. What they didn't know, though, was that another assistant was in the room at the back of the shop. He had caught sight of the two thieves and had pressed a small alarm bell. At that precise moment, a police patrol car was at the junction between Main Street and the road that runs past the library – you know, near the Ritz.

2. X: Write the letter B at the junction referred to by the speaker.

A: So things went wrong for the robbers from the start?

B: Yes. By the time they were leaving the jeweller's, the police car was already turning into Clifton Road. The two men hadn't even time to close one of the car doors properly as they set off in the direction of the football stadium. A passer-by heard one of the men tell the driver to take the first turning off Clifton Road. Just after that, one of the briefcases fell out as their car swung left.

3. X: Where did the thieves lose one of the briefcases? Write the letter C on the spot.

B: But that wasn't the end of it all. When they looked back, they saw the police car gaining on them.

A: But they didn't give up?

B: No, they accelerated. They turned left and then they turned right. Then they swung into a narrow street and stopped a few yards down it at the side of a second car – most likely their getaway car.

4. X: Where was the getaway car parked? Write D on your map.

A: Had they managed to throw off the police car?

B: No. As they were about to change cars, they heard it coming up behind them. So they changed their minds and started off again in the red car. At the end of the narrow street, they turned left into Hilliard Road again and sped off in the direction of the stadium. At the next but one junction before the stadium – you know, on the south side of it – a second police car suddenly cut across their path and forced them to stop.

5. X: Write E at the place where the robbers were forced to stop.

A: What on earth did they do then?

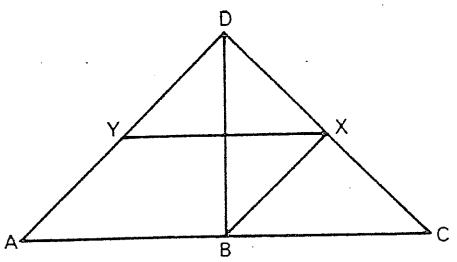
B: Well, by this time, they were really desperate. The driver of the red car got out and fired a pistol at the police car. But this didn't stop the police. One of them scrambled over the bonnet of the police car and chased the man with the pistol down Hilliard Road.

The man ran in the direction of the City Hall and then took the first turning left after Regent Lane. Then he made as if to turn right, but just at that moment, he tripped and fell. In a matter of seconds, two policemen were on him.

6. X: Where was the man caught? Write the letter F to show the place.
   A: Where were the other two robbers while all this was happening?
   B: Well, by this time one was halfway down Regent Lane with two passers-by chasing him. He slipped through the narrow alley at the end of the lane and turned right before dashing across the road. He got most of the way down the road to the station before one of the passers-by finally caught up with him and overpowered him.
7. X: Where was this robber caught? Write the letter G to show the place.
   A: Amazing! I suppose the police soon caught the third man.
   B: No, they didn't. The third robber had a shotgun and he'd sprinted along Regent Lane and into Main Street. He was about to set off running in the direction of the market. Then he caught sight of a butcher's van travelling towards him. He stood quite still in the middle of the road, pointed his gun at the bewildered driver and shouted to him to stop and get out.
8. X: Write the letter H to show where the robber stopped the van.
   B: Then he got into the van, started off down Main Street and turned left only to find himself in the middle of the bus station! He quickly turned round and headed up Main Street. Next he took the road leading to the main entrance of the football stadium. Halfway down this road, however, he saw two policemen on motorcycles in front of him at the end of the road.
9. X: Write the letter I where the two police motorcyclists were.
   A: Well, he must have been well and truly cornered by now.
   B: Yes, but he still fired several shots at the motorcyclists. Then he reversed and jumped out of the van at the end of the road. He turned in the direction of the City Hall. He hadn't got more than a yard or two when he found himself surrounded by a dozen policemen.
10. X: Write the letter J to show where the third robber was caught.
    A: And so at last he was caught!
    B: Yes, and so was the manager of the jewellery shop.
    A: What on earth do you mean?
    B: Well, the police have just found that it wasn't real jewellery at all. It was imitation stuff. So the jeweller's been arrested for fraud!

Type 6 There are many other ways of exploiting visual materials for testing simple listening comprehension. The following kind of item may be useful in the testing of the listening ability of students of mathematics.

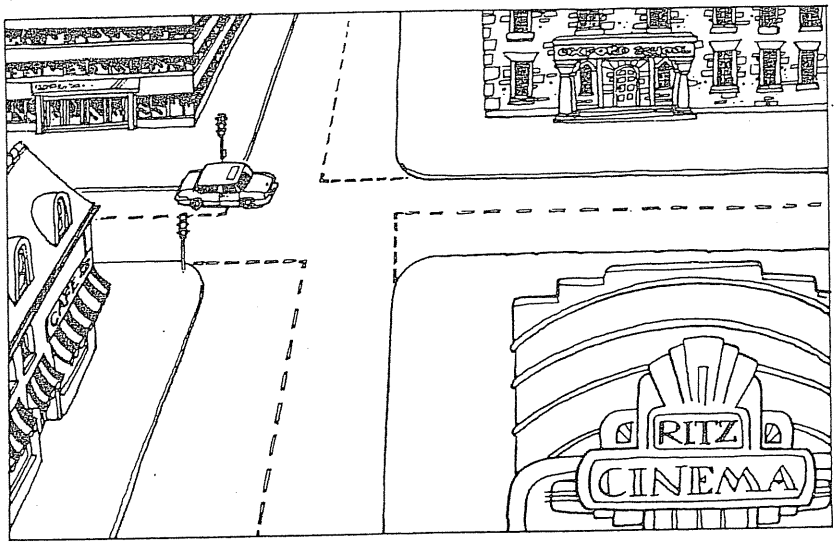
1. Draw a straight line AB three centimetres long.
2. Continue the line AB to point C so that AC is twice as long as AB.
3. Draw a perpendicular from point B.
4. Measure an angle 45° to be called DAC.



5. Now draw the line AD until it meets the perpendicular at point D.
   Question 1: How long is AD?
6. Now join DC.
   Question 2: What does angle ADC measure?
7. Draw a line from point B parallel to AD and mark the point X where it bisects CD.
   Question 3: How long is BX?
8. Now draw a line from X parallel to AC so that it bisects AD at Y.
   Question 4: How long is AY?
   Question 5: How many figures have you drawn?
   What are they?

Type 7 Another useful test item (and exercise) which is independent of the speaking, reading and writing skills is that in which the testees are presented with an incomplete picture (usually a simple line drawing) and are required to add to it pieces of visual information according to certain oral instructions they are given. The following is an example of such an item:

(The testees look at the picture)



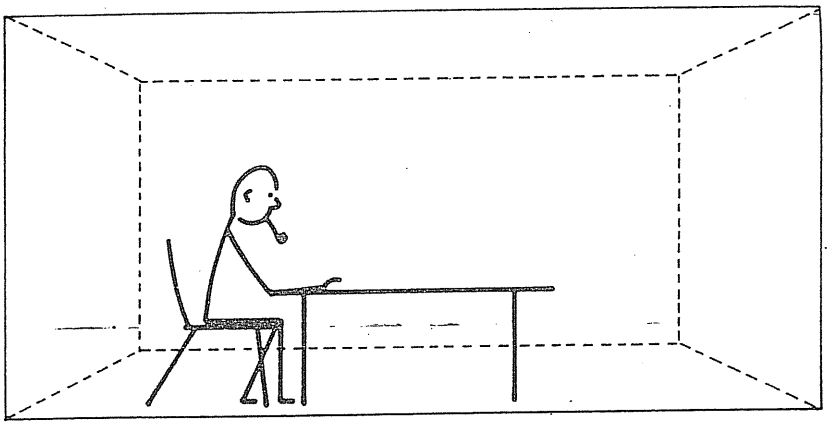
The testees hear:

1. Draw a table and two chairs in front of the café.
2. Draw two traffic lights on the opposite side of the road.

3. Draw a zebra-crossing between the Oxford School and the cinema.
4. Draw a lorry travelling in the opposite direction to the car just before the junction.
5. A policeman directing traffic is standing in the middle of the junction. Draw him.
6. Although there's only one tree at the side of the office building, there are two trees on the opposite side of the road. Draw them.
7. Some people have complained about the danger of crossing the road between the café and the cinema. A pedestrian footbridge has now been built at this point. Draw it.
8. A man who has been cleaning the windows of the second floor of the office building opposite the café has forgotten to take his ladder away. It's still leaning against the window on the extreme right of the front of the building. Draw it.

It is clearly important to keep any such drawing simple so that too much is not required from the testees. Basic practice in matchstick drawings would be a useful preparation for such listening comprehension tasks in class progress tests. A simple country scene involving the drawing of cows, trees, tents, pots and pans, rivers, bridges, fish, birds or an indoor scene involving the positions of furniture and simple objects might form a useful basis for such a listening test. Moreover, it is important to try out this kind of activity before giving it as a test item to students. This kind of pre-testing of items will help to avoid such problems as leaving students insufficient room to draw all the various objects required.

It is also useful to build up an interesting story instead of limiting the comprehension test to instructions in single sentences. The following listening test, constructed by a teacher some time ago, could have been much more interesting had it been put in the form of a simple story or sequence of events.



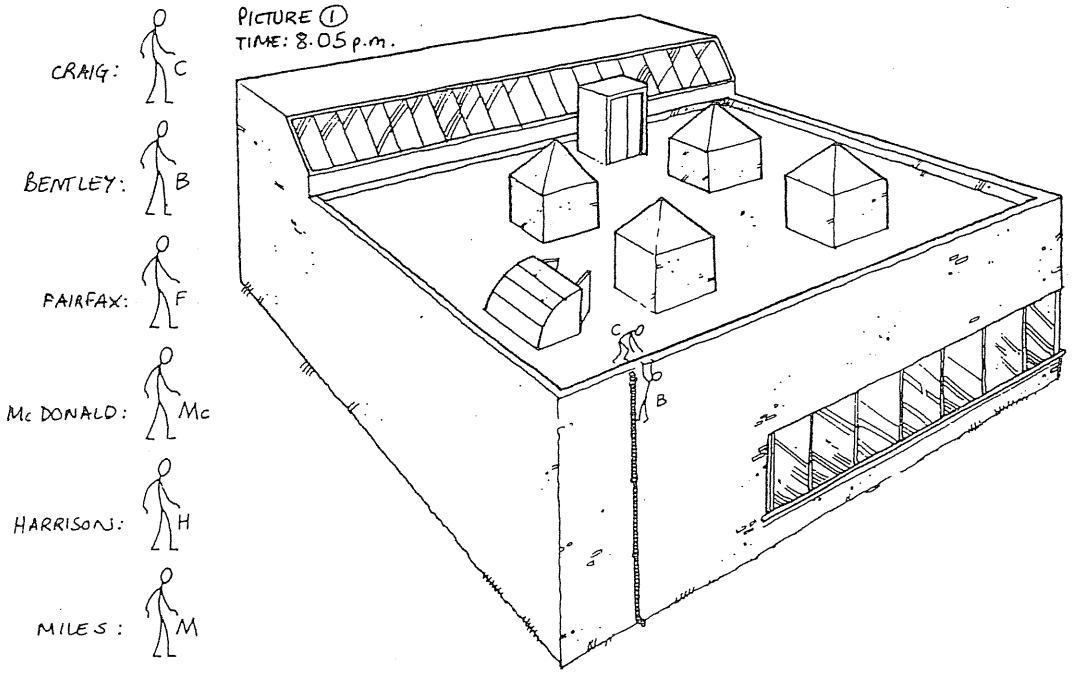
Look at the drawing and listen to the sentences. You will see Mr Peanut sitting at a table. Do what I tell you and complete the picture. It doesn't matter if you cannot draw very well. Are you ready? Now begin.

1. Mrs Peanut is sitting on a chair opposite Mr Peanut. *PUT HER IN.*
2. The door is behind Mr Peanut's back. *PUT IT IN.*
3. Mr Peanut can see Mrs Peanut but he can't see Charlie because Charlie's standing behind him. *PUT HIM IN.* . . (etc.)

The test was later put into the form of a short narrative and found to be far more interesting:

> One day Mr and Mrs Peanut were sitting down to have a meal when something strange happened. Mr Peanut had his back to the door and Mrs Peanut was sitting opposite him. Draw them in your picture. (*Pause*) Their small pet dog was asleep under the table halfway between them – a little nearer Mr Peanut than Mrs Peanut. Draw the pet dog. (*Pause*) Suddenly the door was flung open and a young man rushed into the room and stood behind Mr Peanut, holding a long knife. Can you draw him? (*Pause*) (etc.)

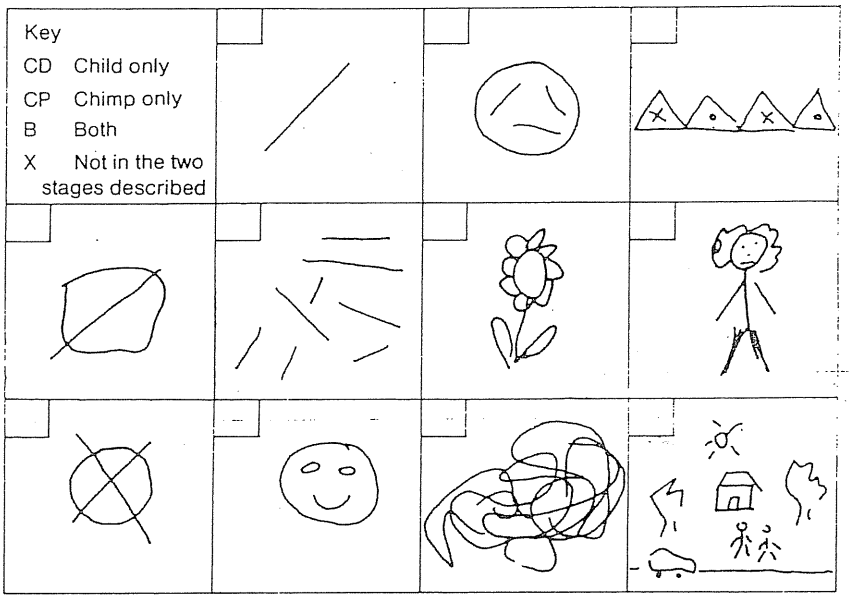
When put into the form of a simple narrative, the test at once becomes interesting. The following diagram is used for a listening test based on the reconstruction of an actual battle on a warehouse roof between two youths and the police.



The test[6] consists of two people discussing what happened, minute by minute, in the events leading up to the shooting of a policeman. Students are asked to plot the positions of the police and the youths on the roof of the warehouse at various times in this gun battle, according to the information they hear on the tape. Several copies of the plan of the warehouse roof are given to each student so that they can draw the positions of the individuals at the various times stated in the discussion. Indeed, it is only through building up a picture of the sequence of events in this way that it becomes possible to gain an understanding of what really happened on the night of the shooting. As a result of attempting to do this, students experience a deep interest and a real sense of purpose when listening to the discussion.

Type 8 Several well-known examining bodies frequently use pictures based on a listening comprehension talk in order to test students' ability to identify and classify information which they have heard. The following is an example[7] of such a type of item.

> In this part you will hear a talk about drawings made by chimpanzees and children. You will hear information twice. Then you should:
> 1. identify the nine sentences, from A to K below, which are about the development of a child. Number them in the order they occur. Write the numbers 1–9 in the boxes.
> 2. classify the diagrams by writing above each diagram an appropriate symbol from the key.



A. shapes inside circles ☐
B. single lines ☐
C. human figures ☐
D. single shapes ☐
E. overlapping shapes ☐
F. regular patterns ☐
G. unconnected lines ☐
H. everyday objects ☐
I. rows of shapes ☐
J. masses of connected lines ☐
K. human faces ☐

Type 9 Simple paper-folding and drawing can also be used to measure ability to understand instructions. For example:

> (*Spoken*) Print your name in the top left-hand corner of your paper: draw a one-inch line six inches under it. Draw a small circle on the line and on the right of it draw a square roughly the same size as the circle. Now take the top right-hand corner and the bottom left-hand corner and fold your paper so that the drawing appears on the outside.

Practical considerations, however, should not be ignored in the administration of this type of test. Although useful for ordinary classroom purposes, such tests are difficult to administer in public examinations. Copying is a simple matter and test reliability may thus be greatly affected.

Moreover, there is a tendency for such tests to become tests of intelligence rather than of language proficiency. The item writer must be careful to test only the student's ability to understand a spoken message – not the ability to interpret it and see hidden relationships.

6.6 Understanding talks and lectures

The ability to understand both informal talks and formal lectures is an important skill for students studying subjects in the medium of English at intermediate and advanced levels.

Type 1 Testees listen to a short talk and select the correct answer about the talk.

Spoken: There's a marked tendency for most developed countries to grow steadily noisier each year. This continually increasing amount of noise is uncomfortable and, what is more important, can affect our health. The noise of machines, heavy traffic and aeroplanes constitutes perhaps the most serious threat to public health. Such noise can interfere with our ability to converse, it can disturb our sleep, and it can quickly make us become nervous wrecks. A loud blast or an explosion may even cause damage to our hearing. But there's another danger – just as great. This is the gradual damage which may be caused if we're continually exposed to noise over several years. Such exposure to noise can undermine our health – as well as our performance and efficiency. Fortunately, technology is progressing at a very rapid rate. Some manufacturers are now designing new silencing mechanisms in their products, and planning experts are even beginning to plan cities according to sound zones.

Written: Only one of the following statements about the talk you have just heard is correct. Put a circle round the letter next to the correct statement.

A. Modern technology is now making towns in developing countries free of loud noise.
B. The increase in noise is a problem which cannot yet be solved by modern technology.
C. Gradual noise over a long period may have just as harmful an effect as loud or sudden noise.
D. There is no real solution to the problem of increasing noise in modern life.

Type 2 Like Type 1 this test combines listening comprehension with reading comprehension. The testees hear a short talk and then read a summary containing blanks. They must then complete the blanks from the talk they have heard. The danger here, however, is that testees could successfully complete the written summary of the talk even if only little had been understood.

Spoken: Would you like a robot in your house? It's now generally accepted that in the future robots will take over many of our tasks, especially jobs of a repetitive nature. But it's highly doubtful if robots will ever be able to do any of the more creative types of work – or indeed if people would want them to. In the home, robots would probably be used to do the cleaning, table-laying, scrubbing and washing-up, but it's considered unlikely so far that they'll be used to do the cooking – at least, not in the near future. According to engineers,

robots will do nothing more original or sophisticated than they have been programmed to do by human beings. And so robots in the home might not be creative enough to do the cooking plan the meals, and so on. They would be used as slaves, thereby freeing people to do more of the things they wanted.

In factories, mobile robots would carry out all the distribution and assembly tasks while human beings carried out research and drew up plans for new products. Human beings would still be responsible for diagnosing faults and for repairing and maintaining machinery. On the farm, robots would probably drive tractors; they'd be programmed to keep their eyes on the ground in front to guide the tractor along a straight line or between rows of vegetables.

The robots themselves would probably not look at all like human beings because their design would be chiefly functional. For instance, it would not be at all surprising to find a robot with an eye in the palm of its hand and a brain in one of its feet!

Written: The following passage is a written summary of the short talk you have ~~just~~ heard. Give the correct word which can be used in place of each number.

In future (1) will do many jobs, particularly those which are (2) by nature. It is generally doubted if they could do (3) work and in the home they would probably not do things like (4). Robots will do nothing more (5) than they have been (6) to do by human beings. A robot would be a kind of (7), freeing human beings so that they could do whatever they wanted. Although robots would be used in factories, human beings would probably (8) the machinery. On farms, robots would probably drive (9). The robots would look (10) human beings because they would be (11) in design. It would even be possible for a robot to have an eye in its hand or a (12) in one of its feet.

Type 3 The testees hear a short talk or lecture and are required to answer questions on it. Unless they are allowed to take notes on the talk, the test may put too heavy a load on the memory. In certain instances, in fact, it may be desirable to give them some guidance for note-taking. The provision of a list of points on which questions will be asked may improve the test.

The following is an example of a test based on a (fictitious) novelist; the questions that follow relate to the novelist's place of birth, early influences on his childhood, the books he read at school, his first publications, his travels, etc. The sheet given to each testee a few minutes *before* the lecture reads as follows:

NOTE PAPER

You are going to hear a talk about Charles Edward Blackwell, a writer of children's books. You are being tested on your ability to listen and understand. After the talk you will be asked 25 questions about Charles Edward Blackwell.

This sheet of paper is for any notes which you wish to take while you are listening to the talk. The notes will not be marked in any way by the examiner.

The questions you will be asked after the talk will be about the points listed below. A space has been left to enable you to write notes for each point.

1. What Blackwell enjoys doing
2. Blackwell's birth
3. His age at the time of the economic depression
4. The books Blackwell read

(etc.)

The testees may take notes during the lecture. They will later receive the following answer sheet.

INSTRUCTIONS

You have just heard a talk about Charles Edward Blackwell, a writer of children's books. You are being tested on your ability to listen and understand. You now have 15 minutes to answer the questions which follow. The 25 questions follow the order of the talk and you should complete each statement with the best answer. Write 'A', 'B', 'C', or 'D' on the line provided at the side of each question. *DO NOT WRITE OUT THE FULL ANSWER.* Here is an example:

This talk is about — Ex. C .....
A. writers of children's books.
B. children's reading.
C. Charles Edward Blackwell.
D. Leeds University.

1. Blackwell enjoys — 1. .....
A. writing books for children.
B. giving lectures for writers.
C. reading books to children.
D. talking about himself.

2. When Blackwell was born, his father was — 2. .....
A. a cricketer.
B. an inn-keeper.
C. a writer.
D. a factory worker.

3. At the time of the great economic depression Blackwell was — 3. .....
A. three years old.
B. five years old.
C. twenty-five years old.
D. thirty years old.

4. When Blackwell was a boy, he read — 4. .....
A. books about child geniuses.
B. Tolstoy's *War and Peace.*
C. stories written for boys of his age.
D. advice about writing for children.

(etc.)

This type of test is generally administered in one of the following ways:

1 The testees receive note paper and take notes while they listen to the lecture. They are then given the question paper (usually consisting of multiple-choice items).

2 The testees receive the question paper first and are given a few minutes to glance through it. They then hear the lecture and work through the

questions. The questions are generally in the form of (a) multiple-choice items, or (b) true/false items, or (c) incomplete sentences. Completion, however, is not usually to be recommended as the testees are faced with the tasks of listening, reading and writing simultaneously – an extremely difficult operation even for native speakers. Even multiple-choice items may cause confusion (especially if not carefully spaced out throughout the lecture), since the testees have to listen while reading carefully through all the options and making their selection. Indeed, if this particular procedure is to be adopted at all, it is perhaps best to use true/false type items since this reduces the amount of reading and the selection to be made.

3 The testees listen to the lecture and then receive the question paper. They read it through and then listen to the lecture given a second time. Although the testees will be listening with a purpose during the second reading of the lecture, the criticisms made previously still apply. Moreover, this test does not approximate as closely to a normal lecturing situation as does method 1.

In writing such tests, it is advisable to give a talk from rough notes or to record a talk and then to work from a transcript of the talk in setting suitable questions. The importance of presenting real speech instead of written prose spoken aloud cannot be emphasised too strongly. This, of course, may be very difficult for non-native speaking teachers, but access to radio broadcasts in English or recordings of talks given by native speakers will be of real help in such cases.

There will inevitably be times when the non-native teacher of English will be forced to use written texts as the basis of a listening comprehension test. In these cases, it is important to keep to the normal delivery rate, increasing the length of pauses at the end of breath segments (i.e. units of meaning such as phrases, clauses and short sentences). In addition, the written text itself should be adapted to assume the features of speech as far as possible: the important points can be restated in various ways and complex sentences can be rewritten in the form of short compound or simple sentences.

It is most inadvisable to attempt to introduce other essential features of spoken discourse when adapting written texts for the purposes of reading aloud. The more the test writer tries to incorporate such features as hesitation, false starts, and ungrammatical sentences, the more artificial the talk will become. In other words, the deliberate introduction of those spontaneous elements inherent in all speech will only increase the artificiality of the situation. In an experiment conducted several years ago, a report about the events in an imaginary East African country was read aloud and recorded for a listening test by a teacher playing the part of a reporter who had just returned from the country in question. The test proved fairly successful even though the students were required at first to listen to written discourse carefully prepared and read aloud. The same report was then read aloud to another group, the speaker deliberately introducing hesitation features, making false starts, etc. Though taking more time to deliver and imparting information at a slower rate, the report was not only extremely difficult to understand but also very laboured and irritating for the listeners. A third presentation of the same report was then given, but this time it took the form of an interview. The 'reporter' began in a self-conscious way, acting his part and speaking well-rehearsed lines.

However, the person conducting the interview became so interested in the subject that he put the script aside and asked questions which the 'reporter' was not expecting. Suddenly the interview came to life and was far more natural and spontaneous. The interview itself lasted much longer than either of the previous talks and new questions were required, but the listening test had at once become valid.

A number of listening tests involve extra-linguistic factors – memory, knowledge of a topic and interest in that topic. It is thus important to avoid testing memorisation of unimportant and irrelevant points in a talk (e.g. *When was the writer's grandfather born?*). There is little justification for setting questions on such points in a reading comprehension test – and far less in a test of listening comprehension. The taking of notes also minimises the memory factor, but the test itself may then become more a test of note-taking skills.

Above all, remember that it is the propositional meaning of sentences which is retained by the listener (i.e. their *general* meaning and intention) and not the actual words or grammatical forms used to express that meaning.[8] For example, the general meaning is still the same whichever of the following two sentences is used:

> The weary band of explorers managed to cross the wide river in a very small yacht.
> The tired party of explorers succeeded in getting to the other bank of the great river in a tiny dinghy.

We are rarely called upon to remember the exact words someone spoke in real life unless in very unusual circumstances, e.g. evidence given in a court case, in which a speaker's exact words may have great significance Even in such circumstances, individuals usually have great difficulty in recalling the actual words spoken even though they can remember perfectly the general meaning of what the person said. Therefore, avoid setting questions which involve the memorisation of individual words in sentences. If a summary of a talk is given for completion, the words omitted in the summary should be those words essential to the meaning of the whole talk (e.g. the word 'robots' in Type 2 in this section).

Type 4 It is important that the content of the text itself should determine the type of question or item used to test comprehension. There is nothing intrinsically either right or wrong about the use of multiple-choice items for a listening comprehension test. It is essential, however, that the most appropriate type of item should be used. Certain texts will lend themselves far more to multiple-choice items than others; other texts will lend themselves to questions set in a tabular form; others will suggest visuals, while yet others are better exploited if followed by open-ended questions.

The following example[9] is included to show how true/false items can be developed to include a third choice (no information available). This is a particularly useful device in listening and reading comprehension tests since frequently the information being sought is not contained in a text. It is just as useful to test the ability to be aware of important information *not* given in a talk as it is to test information given in the talk. Furthermore, the choice becomes no longer a two-way one but a three-way one, thereby reducing the effect of guessing.

	Yes The statement is true.	No The statement is false.	? We have no information about it.
They had a problem taking off because they were carrying so much fuel.			
They started on June 15th, 1919.			
It was foggy when they took off.			
They had a problem when something fell off the plane during the flight.			
They had difficulty finding their way.			
They had no heating in the plane. A snowstorm affected the engine.			
They were injured when they landed.			
There were a lot of problems on the flight.			

Notes and references

1 Such item types have been used in the University of Cambridge Local Examinations Syndicate *Lower Certificate in English* but are now no longer used in the *First Certificate in English* examination.

2 Elisabeth Ingram reports that native speakers of American English made up to 10 per cent errors in the ELBA phoneme sub-test as compared with an average of 2 per cent errors made by native British speakers (*Language Testing Symposium*, Oxford University Press 1968).

3 This item type was devised by Elisabeth Ingram for use in the *ELBA Test* (English Language Battery).

4 Robert Lado made use of this technique in his *Test of Aural Comprehension*.

5 A version of this test item used to be included in the *Graded Achievement Tests in English* (GATE) ~~of the~~ American Language Institute, Georgetown University.

6 The test item was written by David Bonamy and John Beverley.

7 Joint Matriculation Board, *Test in English (Overseas)*, Oral Paper, June 1983.

8 For further details, see Richards, J C 1985 *The Context of Language Teaching*. Cambridge University Press

9 The Royal Society of Arts: *Examinations in the Communicative Use of English as a Foreign Language*, Intermediate level, Test of Listening, May 1983.

7

Oral production tests

7.1 Some difficulties in testing the speaking skills

Testing the ability to speak is a most important aspect of language testing. However, at all stages beyond the elementary levels of mimicry and repetition it is an extremely difficult skill to test, as it is far too complex a skill to permit any reliable analysis to be made for the purpose of objective testing. Questions relating to the criteria for measuring the speaking skills and to the weighting given to such components as correct pronunciation remain largely unanswered. It is possible for people to produce practically all the correct sounds but still be unable to communicate their ideas appropriately and effectively. On the other hand, people can make numerous errors in both phonology and syntax and yet succeed in expressing themselves fairly clearly. Furthermore, success in communication often depends as much on the listener as on the speaker: a particular listener may have a better ability to decode the foreign speaker's message or may share a common nexus of ideas with him or her, thereby making communication simpler. Two native speakers will not always, therefore, experience the same degree of difficulty in understanding the foreign speaker.

In many tests of oral production it is neither possible nor desirable to separate the speaking skills from the listening skills. Clearly, in normal speech situations the two skills are interdependent. It is impossible to hold any meaningful conversation without understanding what is being said and without making oneself understood at the same time. However, this very interdependence of the speaking and listening skills increases the difficulty of any serious attempt to analyse precisely what is being tested at any one time. Moreover, since the spoken language is transient, it is impossible without a tape recorder to apply such procedures as in the marking of compositions, where examiners are able to check back and make an assessment at leisure. The examiner of an oral production test is working under great pressure all the time, making subjective judgements as quickly as possible. Even though samples of speech can be recorded during a test, the tape-recording, by itself, is inadequate to provide an accurate means of reassessing or checking a score, since it cannot recapture the full context of the actual situation, all of which is so essential to any assessment of the communication that takes place.

Yet another (though not insuperable) difficulty in oral testing is that of administration. It is frequently impossible to test large numbers of students because of the limited time involved. It is not difficult to appreciate the

huge problems relating to a test situation in which thousands of students have to be examined by a handful of examiners, each student being tested for a period of, say, ten or fifteen minutes. Although the use of language laboratories for such tests has made it possible in some cases to administer more reliable oral production tests to large numbers of students, the actual scoring of the tests has not been so easily solved.

Excluding tests of reading aloud and one or two other similar tests, oral tests can have an excellent backwash effect on the teaching that takes place prior to the tests. For example, in one country the oral test was retained as part of a school-leaving examination simply to ensure that at least some English would be spoken in the last two years of the secondary school – even though the test itself was considered an unreliable measuring instrument as a result of the large number of unqualified examiners who had to administer it. For this reason, and indeed, because oral communication is generally rated so highly in language learning, the testing of oral production usually forms an important part of many language testing programmes.

The following sections in this chapter will give an idea of the range of possible types of oral tests. Some of the exercises (e.g. picture descriptions) have proved very useful in many tests while others (e.g. pencil-and-paper tests) have met with varying degrees of success. In spite of its high subjectivity, an extremely good test is the oral interview. In many cases, one or two sub-tests (or oral activities) are used together with the oral interview to form a comprehensive test of oral production skills.

7.2 Reading aloud

Many present-day oral tests include a test of reading aloud in which the student is given a short time to glance through an extract before being required to read it aloud. The ability to read aloud differs greatly from the ability to converse with another person in a flexible, informal way. Although reading aloud may have a certain usefulness, only a few newsreaders and teachers may ever require training and testing in this particular skill. The majority of students will never be called on to read aloud when they have left school. It is a pity, therefore, that students are required to sacrifice their enjoyment of silent reading in order to practise reading aloud. We read primarily for information or enjoyment, and the silent reading skills so necessary for this purpose differ greatly from those of reading aloud. The backwash effects of this kind of test may be very harmful, especially in areas where the reading skills are misguidedly practised through reading aloud. Finally, how many native speakers can read aloud without making any errors?

Tests involving reading aloud are generally used when it is desired to assess pronunciation as distinct from the total speaking skills. In order to construct suitable tests of reading aloud, it is helpful to imagine actual situations in real life in which the testees may be required to read aloud. Perhaps one of the most common tasks is that of reading aloud directions or instructions to a friend, colleague or fellow-worker: e.g. how to wire a plug, how to trace faults in a car engine, how to cook certain dishes. For example, the following instructions relate to a situation in which a teacher or class monitor may be asked to read aloud:

> First put the headset on. Make sure it is in its most comfortable position with the headband over the centre of the head. The microphone should be about 1½ inches from the mouth.

To record, put the white switch to the position marked *Work*. Put the red switch to *Speak* and press the red recording button, which will now light up.
(etc.)

Another situation which might occur in real life is that in which the student is asked to read aloud (part of) a letter he has received. For all the different extracts, however, it is advisable to draw up certain features which must be included in each passage: e.g. one Yes/No question, one Wh- question, two sentences each containing a subordinate clause, one question tag, the phoneme contrasts i:–ɪ, p–b, ɒ–ɔ:, etc. In this way, some degree of consistency can be achieved.

A test more useful in many ways than reading aloud is the retelling of a short story or incident. In this type of examination, the students are required to retell a story they have just read. If carefully constructed, such a test can assess most of the phonological elements which are otherwise tested by reading aloud. Unfortunately, it often measures other skills such as reading comprehension, memory and organisation, too.

7.3 Conversational exchanges

These drills are especially suitable for the language laboratory and can serve to focus attention on certain aspects of the spoken language, especially in those countries where English is taught as a foreign language and the emphasis is primarily on the reading skills. However, several of the test items themselves are far from communicative in any sense at all and do not allow for authentic interaction of any kind. The essential element of constructive interplay with unpredictable stimuli and reponses is absent from all these items as a result of the attempt to control the interaction taking place. The item types range from items presenting the testees with situations in which they initiate conversations to incomplete conversations with the part of one speaker omitted (i.e.a one-sided dialogue). Tests containing such item types are on the whole reliable, but they cannot be described as being valid tests of speaking. If an opportunity is provided in other parts of the test for real oral interaction (i.e. genuine conversation and discussion), however, these controlled test items can be of some use in directing the attention of the students to specific language areas and skills.

Type 1 The testees are given a series of situations and are required to construct sentences on the lines of a certain pattern or group of patterns. Again, it is essential that two or three models be given to the testees so that they know exactly what is required. (The testees read or hear the situation and then make the appropriate responses, shown in the brackets.)

Examples:

Mrs Green lives in a flat. She doesn't like living in a flat and would like to live in a small house with a garden. (*She wishes she lived in a small house with a garden.*)

It's raining heavily. Tom and Anna are waiting impatiently at home to set off on their picnic. (*They wish it would stop raining.*)

1. Mr Black has a small car but his neighbours all have large cars. He would like a large car, too.
2. Anna hasn't learnt how to swim yet but most of her friends can swim.
3. Tom is waiting for Bill outside the cinema. The show is just about to start but Bill has not arrived yet.

4. Mrs Robinson doesn't like living in towns; she wants to live in the country.

(etc.)

Type 2 This type of test item is similar to the previous type but not as strictly controlled.[1] No model-responses are given by the examiner and the students are free to use whatever patterns they wish.

A friend of yours has forgotten where he has put his glasses. He cannot see too well without them. What will you say to him? (*Let me help you to look for them*, etc.)

You are on your way to school when it starts to rain heavily. Unfortunately, you and your friend have no raincoats. There is nowhere to shelter but your school is only a hundred yards away. What do you say to your friend? (*Shall we make a dash for it?/Let's run the rest of the way.*)

1. You are trying to get to the public library but you are lost. Ask a police officer the way.
2. Your friend has just returned from a holiday abroad. What do you say to him?
3. A waitress has just brought you the bill but has totalled it up incorrectly. What do you say to her?
4. A friend of yours wants to see a film about a murder. You have already arranged to see it another evening, but you know she would be hurt if she knew. Make up an excuse.

Type 3 The students hear a stimulus to which they must respond in any appropriate way.[1] (This test often relies on conventional greetings, apologies, acceptable ways of expressing polite disagreement, etc.)

Do you mind if I use your pencil for a moment?
(*Not at all/Certainly/Please do/Go ahead*, etc.)

What about a game of tennis?
(*Yes, I'd love a game/All right. I don't mind/Don't you think it's a bit too hot?*, etc.)

1. Please don't go to a lot of trouble on my behalf.
2. Oh dear, it's raining again. I hope it stops soon.
3. We shan't be late, shall we?
4. Karen asked me to say she's sorry she can't come tonight.

Type 4 This is similar to the previous type of item, but the stimuli and responses form part of a longer dialogue and the situation is thus developed. Because of its total predictability, however, this type of item is sometimes referred to as a dialogue of the deaf! The man in the dialogue below continues regardless of what the testee says.

You're on your way to the supermarket. A man comes up and speaks to you.

MAN: Excuse me. I wonder if you can help me at all. I'm looking for a chemist's.
PAUSE FOR TESTEE'S REPLY
MAN: Thank you. Do you know what time it opens?
PAUSE FOR TESTEE'S REPLY
MAN: Thanks a lot. Oh, er, by the way, is there a phone box near here?

PAUSE FOR TESTEE'S REPLY

MAN: Oh dear, I'll need some coins. Do you have any change for a £5 note?

PAUSE FOR TESTEE'S REPLY

MAN: Well, thanks a lot. You've been most helpful.

This dialogue clearly becomes absurd if, when asked where there is a chemist's, the testee replies, 'I'm sorry, I don't know,' and the man promptly thanks him and asks what time it opens. Nevertheless, the use of pre-recorded material of this kind makes it possible to use the language laboratory to test large numbers of students in a very short time.

Type 5 This item[2] takes the form of an incomplete dialogue with prompts (shown in brackets in the following example) whispered in the student's ear.

You are at the reception desk of a large hotel. The receptionist turns to address you:

RECEPTIONIST: Can I help you?
(You want to know if there is a single room available.)

YOU: .....

RECEPTIONIST: Yes, we have a single room with an attached bathroom.
(Ask the price.)

YOU: .....

RECEPTIONIST: Thirty-four pounds fifty a night.
(You want to know if this includes breakfast.)

YOU: .....

RECEPTIONIST: Yes, that's with continental breakfast.
(You have no idea what 'continental breakfast' is.)

YOU: .....

RECEPTIONIST: It's fruit juice, coffee or tea and bread rolls.
(The receptionist is speaking too quickly. What do you say?)

YOU: .....

RECEPTIONIST: Fruit juice, coffee or tea, and bread rolls.
(Book the room for two nights.)

YOU: .....

RECEPTIONIST: Certainly. Room 216. The porter will take your bag and show you where it is.
(Thank the receptionist.)

YOU: .....

7.4 Using pictures for assessing oral production

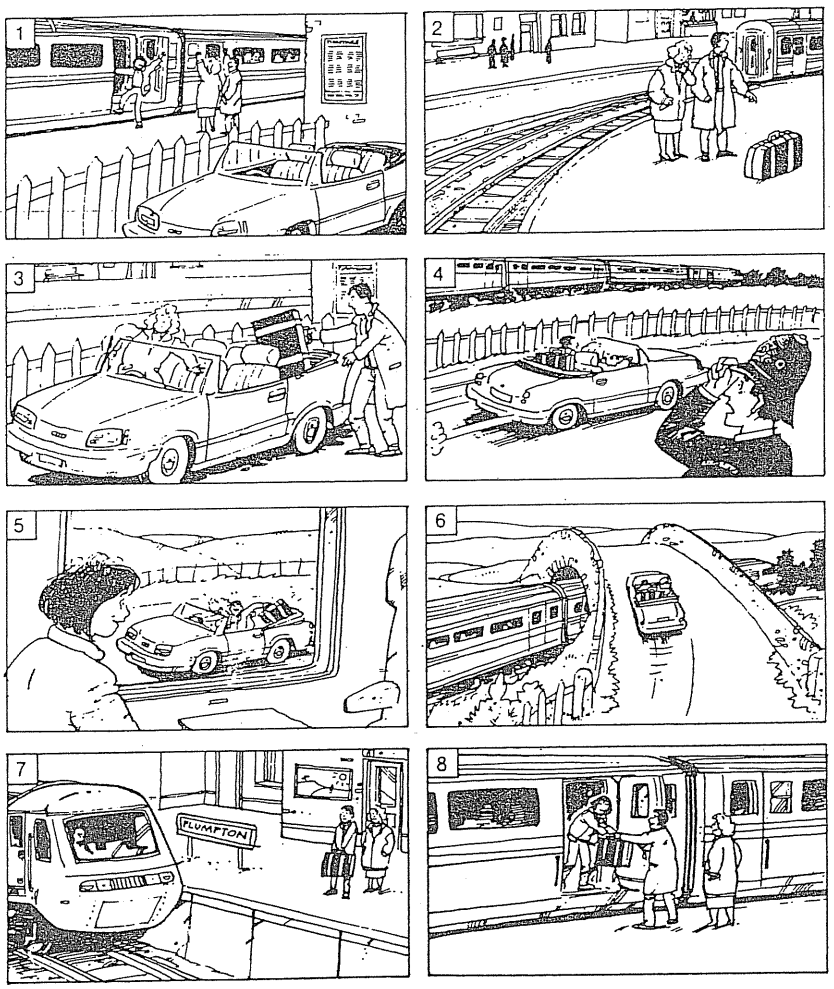
Pictures, maps and diagrams can be used in oral production tests in similar ways to those described in the previous chapter on testing the listening skills. Pictures of single objects can be used for testing the production of significant phoneme contrasts, while a picture of a scene or an incident can be used for examining the total oral skills. This section will concentrate on the use of pictures for description and narration.

The students are given a picture to study for a few minutes: they are then required to describe the picture in a given time (e.g. two or three minutes). Occasionally, the number of words each student speaks is counted by one examiner in the room, while the other examiner counts the number of errors made. The score is thus obtained on the basis of the number of words spoken and the errors made (but this procedure is very unreliable.) Separate scores for general fluency, grammar, vocabulary,

phonology, and accuracy of description/narration are far better. Advertisements, posters and strip cartoons may be used in this way for class tests, provided that there are enough available to prevent the students from preparing one or two set pieces.

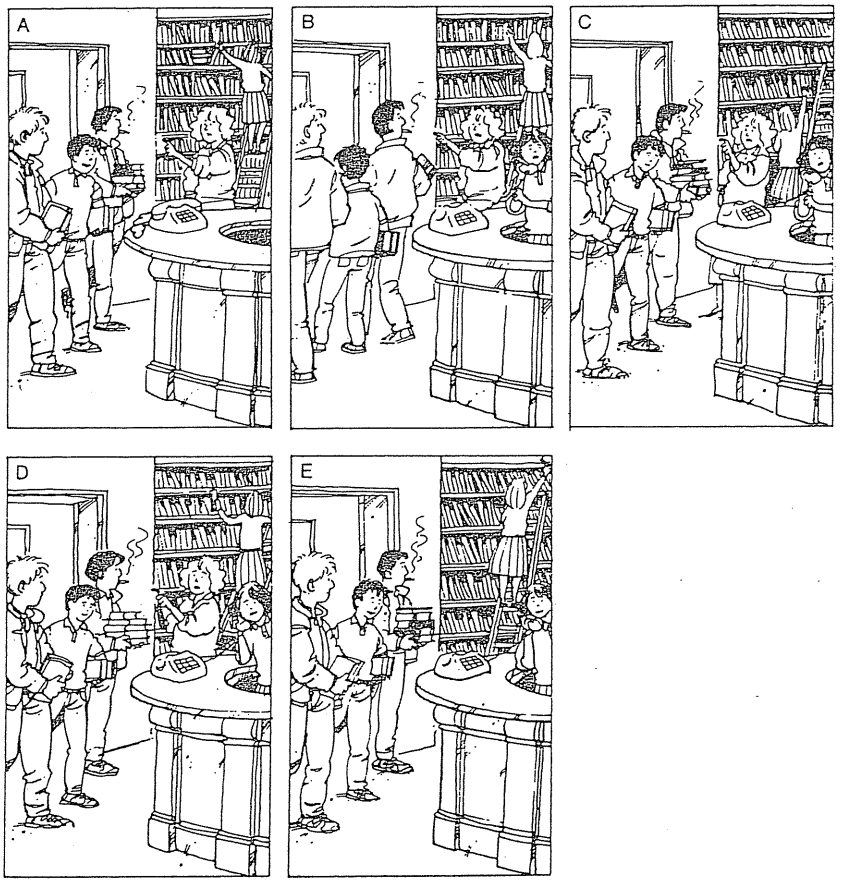
Careful selection of the pictures used for the examination will help in controlling the basic vocabulary required and may, to some extent, determine the type of sentence structure that predominates. Different styles and registers can be tested by including maps and diagrams as well as pictures for comparison, pictures for instructions and pictures for description and narration. If the pictures depict a story or sequence of events, it is useful to give the testees one or two sentences as a 'starter', thereby familiarising them with the tense sequencing they should employ.

> Examiner: Last summer Lucy spent a few days with her uncle and aunt in the country. When it was time for her to return home, her uncle and aunt took her to the station. Lucy had made a lot of friends and she felt sad on leaving them. She got on the train and waved goodbye to them. . . . Now you continue to tell this story.



The most effective type of oral examination using pictures requires not only narration or picture description on the part of the students but also a discussion about the picture(s) concerned. If the examiner asks questions and discusses the picture(s) with each student, the formal speech situation is combined with the reciprocal speech situation and two different types of oral production skills can thus be measured. Even if no discussion is included in the examination, the examiner would be well advised to prompt the student whenever he or she appears to need encouragement. It is always important to find out what a student knows – not what he or she doesn't know: long periods of silence will tell the examiner very little.

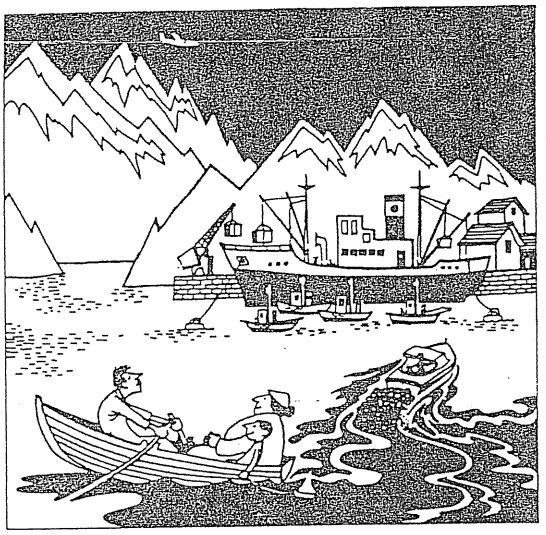
A similar technique[3] to that described in the previous chapter can be used to test oral production. The student and the examiner have five pictures in front of them, each picture differing in only one respect from the other four pictures. The student is given a card bearing a letter (A, B, C, D or E); the examiner cannot see the letter. The student is required to describe the appropriate picture (according to the letter). The examiner then selects a picture according to the description, assessing the student not only on the correctness and fluency of his or her speech but also on the length of time taken before the student's description results in the identification of the appropriate picture. The examiner then checks the card.



Another effective way of assessing a student's ability to speak, however, is to give pairs or groups of students a simple task to perform. Working in pairs, students can describe their own picture before listening to their partner's description of a similar (but not identical) picture. The two students cannot see each other's picture. They can then discuss in which ways the two pictures are the same and in which ways they differ. The pictures used for such discussion may range from almost identical ones to quite different ones. The most successful pictures for this activity, however, will generally depict fairly similar subjects but differ considerably in their treatment of the subject. The following two pictures of a harbour have been used very successfully to generate discussion between two groups of students.



A

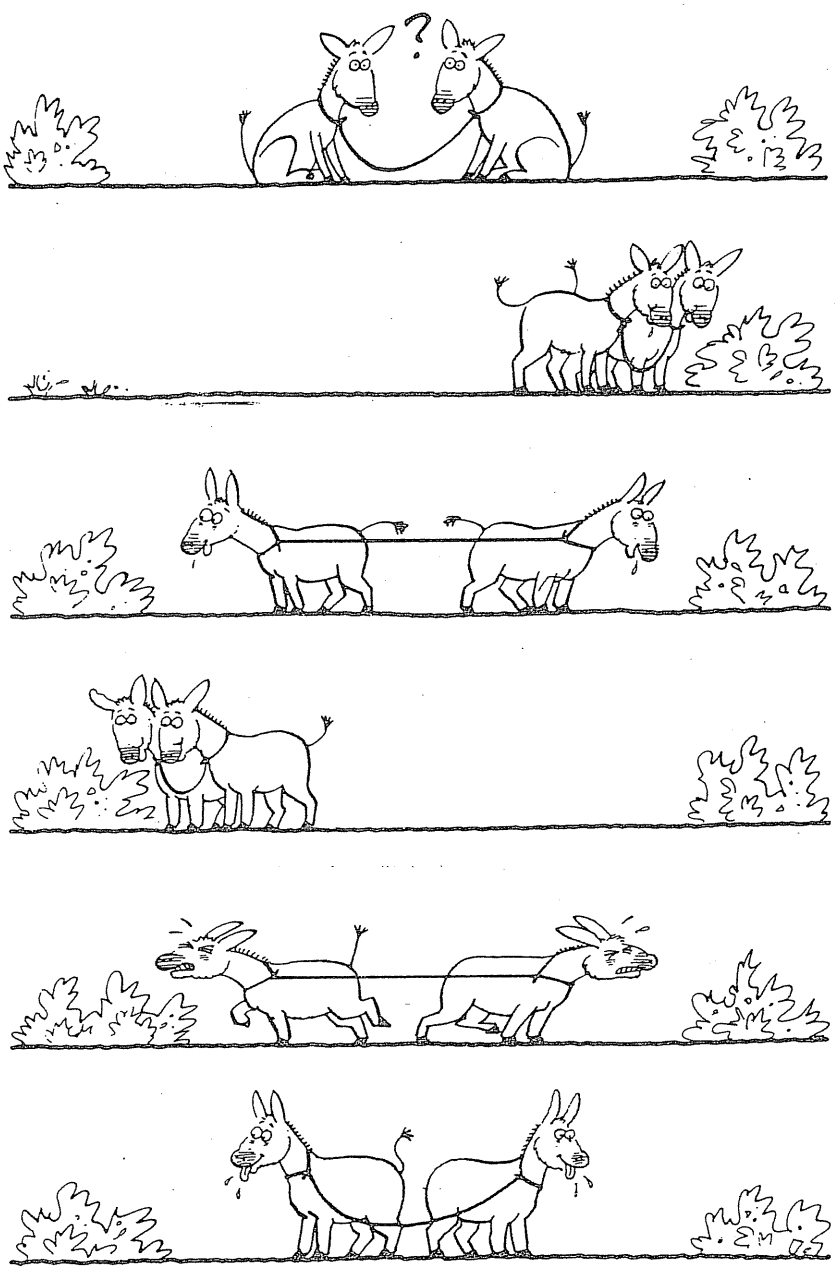


B

When scoring students' performances, the examiner should concentrate on what individual students are doing with the target language and how they are using it to achieve their purpose. Language errors which interfere with successful communication will thus be penalised heavily. On the other hand, those minor errors which, though annoying in certain respects, do not seem to impede communication to any degree will not be penalised in the same way. Clearly, the successful outcome of the activity will thus be important for both examiner and students.

If students are examined in small groups, one of the most useful activities involving pictures is for them to be given a sequence of pictures to rearrange. Students should begin by describing their own picture without showing it to the other members of the group. After each picture has been described and discussed in relation to the other pictures, the group decides on an appropriate sequence. Each member of the group then puts down his or her picture in the order decided upon. Comments are made and the order changed if the group considers it desirable. Again, students are using language to achieve a certain purpose in this type of test.

The following example[4] shows how simple it is to base such an activity on pictures already published, especially those following the comic strip principle.



Two mules

7.5 The oral interview

Like many other examinations of oral production, the scoring of the oral interview is highly subjective and thus sometimes has only low reliability. In addition, the performance of a student in a particular interview may not accurately reflect his or her true ability.

Supporters of the oral interview claim that the examination at least appears to offer a realistic means of assessing the total oral skill in a 'natural' speech situation. Others, however, argue that the examination nevertheless is artificial and unrealistic: students are placed not in natural, real-life speech situations but in examination situations. They are thus susceptible to psychological tensions and also to constraints of style and register necessary in such a situation. For example, many students adopt a quiet and colourless tone in interviews; some even develop a guarded attitude, while others become over-friendly.

One solution to this problem is to have the class teacher as the interviewer: if an external examiner is required, he or she may sit at the back of the room or in any other obscure place. The interviewer (whether teacher or examiner) should endeavour to put the student at ease at the beginning of the interview, adopting a sympathetic attitude and trying to hold a genuine conversation (constantly making his or her own contribution without, at the same time, talking too much). The interviewer should *never* attempt to note down marks or comments while the student is still engaged in the interview. The dual role (i.e. of both language partner and assessor) which the examiner is required to assume in the oral interview is always a most difficult one.

Another solution to problems caused by tension and language constraints is to interview students in pairs or even threes, thus not only putting them more at ease through the presence of a friend or classmate but also enabling them to speak to each other as members of the same peer group. Consequently, the whole atmosphere will become more relaxed and the constraints of register will disappear, resulting in less artificial and stilted language being used. Students will use the language which they normally use in most speech situations in everyday life. No longer will an inferior (i.e. the student) be required to address a superior (i.e. the teacher) throughout the entire interview. Although the teacher, or examiner, may interrupt or direct the discussion whenever necessary, he or she will also be able to adopt a more passive role in the discussion. In this way, students will feel free to converse and use language in a more natural and purposeful way.

The chief danger in conducting interviews with pairs of students is that resulting from personality conflicts or the dominance of one of the members of the pair. It is therefore very important for the teacher or examiner to ensure that the two students forming a pair have similar or sympathetic personalities and have similar levels of language ability. The language task itself can also help since, if the two students are presented with a real problem to solve, there will be a greater chance of them co-operating and working together in the target language. In addition to the use of pictures for comparison and contrast, students can be given simple puzzles and problem-solving tasks. A short quiz may even be prepared by one of the students and given to the other. In all these cases, however, the examiner should be concerned more with the students' use of the target language to achieve their goals rather than with their knowledge or their actual ability to accomplish the task given. Thus, a student's inability to answer several questions in a quiz should not weigh too heavily provided that he or she is able to give some answers (right or wrong) and conduct a reasonable conversation, etc. with the student giving the quiz. In short, each student should be assessed not only on such features of the spoken language as grammatical acceptability and pronunciation, but also on

appropriacy of language and effectiveness of communication – and, where appropriate, the time taken to accomplish the task given.

The oral interview should be scored only after the student has left the room (unless two or more examiners are present – in which case one of them can sit behind the student and score). Although this settles the problem of *when* the interview should be scored, the question of *how* it should be scored still remains. For example, how should the replies to this question be scored?

(Tester) 'What are you going to do this weekend?'
(Student A) 'I'm quite well, thank you.'
(Student B) 'I go to fish. I fish in river near the big wood.

A's reply is perfectly correct but it is nevertheless quite inappropriate: the student simply hasn't heard or understood the question correctly. On the other hand, B's reply shows a real attempt to answer the question but unfortunately contains several errors.

The scoring of the interview can range from an impression mark to a mark arrived at on the basis of a fairly detailed marking scheme (showing accuracy of pronunciation, grammar, vocabulary, appropriacy, fluency and ease of speech). The following marking scheme[5] (using a 6-point scale) is given as just one example of a number of such schemes in present-day use.

Rating	Ability to communicate orally
6	Excellent: on a par with an educated native speaker. Completely at ease in his use of English on all topics discussed.
5	Very good: although he cannot be mistaken for a native speaker, he expresses himself quite clearly. He experiences little difficulty in understanding English, and there is no strain at all in communicating with him.
4	Satisfactory verbal communication causing little difficulty for native speakers. He makes a limited number of errors of grammar, lexis and pronunciation but he is still at ease in communicating on everyday subjects. He may have to correct himself and repattern his utterance on occasions, but there is little difficulty in understanding him.
3	Although verbal communication is usually fairly satisfactory, the native speaker may occasionally experience some difficulty in communicating with him. Repetition, re-phrasing and re-patterning are sometimes necessary; ordinary native speakers might find it difficult to communicate.
2	Much difficulty experienced by native speakers unaccustomed to 'foreign' English. His own understanding is severely limited, but communication on everday topics is possible. Large number of errors of phonology, grammar and lexis.
1	Extreme difficulty in communication on any subject. Failure to understand adequately and to make himself understood.

Note that an even-numbered scale is often preferred because it helps examiners to avoid awarding the middle mark (a tendency in many cases). Thus, although marks may cluster round the median 3 on a 5-point scale, examiners using a 6-point scale will have to decide whether to award 3 (just *below* the middle point on the scale) or 4 (just *above* the middle point on the same scale). It is also advisable to avoid a narrow scale (i.e. a 4-point scale) as this will not allow for the range of discriminations you may wish to make. On the other hand, a wide scale (e.g. a 20-point scale) is not recommended as few markers seem to make use of either the upper or lower ends of this scale, most scores tending to cluster around 9–12.

Many examining bodies prefer fairly short descriptions of grades in order to enable the examiner to glance quickly through the marking scheme. Examiners who are faced with a lot of reading in the assessment will be tempted to rely solely on the numerical grade itself (e.g. 5) – which makes their scoring extremely subjective and liable to fluctuation. Wherever possible, it is useful in public examinations to have two or more examiners listening for particular areas or features before later pooling their assessments. The importance of examiners' meetings and 'practice' interviews (using tape recorders and marking guides) cannot be too greatly emphasised. Such sessions, when conducted with a large number of examiners, are of considerable help in increasing the marker reliability of oral interviews. Sample recordings and scores are discussed and some degree of standardisation of marking is thus achieved.

The previous paragraph largely concerns public examinations or achievement tests set outside the classroom. For most classroom and school tests, however, the teacher should devise his or her own rating scale after having carefully considered the level and kind of oral skills the students should be expected to achieve. The optimum performance expected, therefore, will not necessarily be near native-speaker fluency. Clearly, students who have been spending a few hours a week learning English over a period of two years cannot possibly be expected to achieve band 6 on the scale included here: hence the scale itself would be largely inappropriate for the purposes of most classroom tests of oral English.

In order to devise a suitable scale, the teacher should first begin to describe clearly the criteria for assessing oral ability. The teacher may, for example, wish to consider each student's achievement in terms of accuracy, appropriacy and fluency;[6] accuracy, appropriacy, range, flexibility and size;[7] or fluency, comprehensibility, amount of communication, quality of communication, and effort to communicate.[8] Whatever the criteria selected, the teacher should begin by describing in one or two sentences exactly what he or she expects the average successful student to have achieved under each of the headings by the time the test is taken. These descriptions will then form band 4 (if a 6-point scale is being used). The teacher should then repeat the same procedure for the student who is slightly below average in his or her expectations (i.e. band 3). Next, the teacher writes the descriptions for the successful student (above average), assigning these descriptions to band 5 on the scale. After this, the teacher describes the performance of the unsuccessful student who is below average (band 2), then the performance of the very successful student (band 6), and finally that of the least able and most unsuccessful student (band 1).

The following is an example of a teacher's rating scale for the lower intermediate level.

Accuracy	Fluency	Comprehensibility
6 Pronunciation is only very slightly influenced by the mother-tongue. Two or three minor grammatical and lexical errors.	Speaks without too great an effort with a fairly wide range of expression. Searches for words occasionally but only one or two unnatural pauses.	Easy for the listener to understand the speaker's intention and general meaning. Very few interruptions or clarifications required.
5 Pronunciation is slightly influenced by the mother-tongue. A few minor grammatical and lexical errors but most utterances are correct.	Has to make an effort at times to search for words. Nevertheless, smooth delivery on the whole and only a few unnatural pauses.	The speaker's intention and general meaning are fairly clear. A few interruptions by the listener for the sake of clarification are necessary.
4 Pronunciation is still moderately influenced by the mother-tongue but no serious phonological errors. A few grammatical and lexical errors but only one or two major errors causing confusion.	Although he has to make an effort and search for words, there are not too many unnatural pauses. Fairly smooth delivery mostly. Occasionally fragmentary but succeeds in conveying the general meaning. Fair range of expression.	Most of what the speaker says is easy to follow. His intention is always clear but several interruptions are necessary to help him to convey the message or to seek clarification.
3 Pronunciation is influenced by the mother-tongue but only a few serious phonological errors. Several grammatical and lexical errors, some of which cause confusion.	Has to make an effort for much of the time. Often has to search for the desired meaning. Rather halting delivery and fragmentary. Range of expression often l nited.	The listener can understand a lot of what is said, but he must constantly seek clarification. Cannot understand many of the speaker's more complex or longer sentences.
2 Pronunciation seriously influenced by the mother-tongue with errors causing a breakdown in communication. Many 'basic' grammatical and lexical errors.	Long pauses while he searches for the desired meaning. Frequently fragmentary and halting delivery. Almost gives up making the effort at times. Limited range of expression.	Only small bits (usually short sentences and phrases) can be understood – and then with considerable effort by someone who is used to listening to the speaker.
1 Serious pronunciation errors as well as many 'basic' grammatical and lexical errors. No evidence of having mastered any of the language skills and areas practised in the course.	Full of long and unnatural pauses. Very halting and fragmentary delivery. At times gives up making the effort. Very limited range of expression.	Hardly anything of what is said can be understood. Even when the listener makes a great effort or interrupts, the speaker is unable to clarify anything he seems to have said.

Whatever the criteria chosen, the brief descriptions can be made much more specific at each level in order to reflect the contents of the course being followed. The most important point to bear in mind, however, is that for most classroom purposes the rating scale should not have native-speaker performance as the desired goal. Instead, it should be based on realistic expectations of what successful learners can achieve at a particular stage in their development.

Finally, oral interviews do not simply happen spontaneously. Although each oral interview should simulate as natural and realistic a

speech situation as possible, it is essential that a certain amount of material be prepared beforehand. The interview may be carefully structured or may be structured quite loosely: in both cases, the examiner should have plenty of material on which to fall back if necessary. He or she may lead in to the interview by asking a Yes/No question, followed at some stage or other by certain Wh-questions and question tags.

There are dangers, however, in adhering to a very rigid structure or plan. For example, a student may develop a certain topic and be proceeding happily in one direction when the examiner interrupts and stops the whole flow of the interview in order to include a *How often* question. Again, the demand for certain set responses may reach an absurd stage: in a certain oral examination, for example, it was agreed that *Whereabouts . . .?* ought to signal a different response to *Where . . .?* Consequently, the more helpful students, who anticipated the interviewer, 'failed' on this item:
e.g.

INTERVIEWER: Whereabouts do you live?
(Requiring the name of the *district* where the testee lived.)
STUDENT: 341 King's Road, North Point.

Had the student replied *North Point* in a natural speech situation, the other person would probably have asked: *Where(abouts) in North Point?* In the interview the student merely anticipated such a reaction (consciously or subconsciously) and gave a full answer initially.

Provided that flexibility can be retained, it is useful to prepare a series of questions on a wide range of topics. The following list of 20 topics is given here to help, not to inhibit, the examiner:

family, home, school, sports, hobbies, books, films, transport, week-ends, holidays, radio, health, teeth, shopping, traffic, crime, friends, money, fines, careers, etc.

At least ten or twelve questions can be asked on each topic, but the examiner should never attempt to 'work through' lists of questions. Indeed, the examiner should contribute to the interview from time to time. Examples of the types of questions which can be asked on one of the topics are:

Sports and games
Do you play any games?
What's your favourite sport?
How often do you play?
Are you in your school team?
Do you like watching sport?
Which do you think is the most interesting sport to watch?
Can you swim?
How did you learn to swim?
Whereabouts do you go swimming?
Which game would you advise me to take up? Why?
Which is the most difficult game to play?
Which is better as an exercise: basketball or football?

Where considered necessary, current affairs and highly controversial issues may be introduced in an interview to stimulate or provoke a student, provided that some allowance is made for the emotive content of the

discussion. It is even more important here that the examiner should remain flexible and vary the range of questions for discussion.

If the oral interview is recorded on tape, the examiner can score the interview at leisure, playing and replaying sections where necessary. Extensive notes can be taken from the recordings, comparisons made and confirmation sought where there is doubt concerning a particular mark. If interviews are recorded, the examiner must of course take care to identify each student at the beginning of the interview, especially when comparisons involving rewinding and replaying are to be made.

7.6 Some other techniques for oral examining

The short talk

In certain examinations students are required to prepare a short talk on a given topic. They may be allowed several days or only a few minutes in which to prepare the talk and, in some cases, they may be provided with notes or reference material. This is clearly a realistic test of sustained speech but it constitutes an extremely difficult examination for second-language learners at all but the most advanced stages. Indeed, this particular type of examination is generally very difficult for first-language speakers. The examination can be improved slightly by reducing the time allotted for the talk and asking students questions based on their talk, thus introducing a reciprocal speech situation. The questions might be asked either by the examiner or by a group of students (if the talk is given in front of an audience). In whatever situation the talk is given, however, the examiner must make every attempt to put the students at ease.

Care must be taken to prevent students from learning whole sections of their talk. Subjects about which an individual student knows very little should be avoided. Experience of such examinations has shown that candidates talk better when they have something worthwhile to say and can bring into the talk a genuine interest in the subject coupled with experience and imagination. A co-operative audience also helps greatly.

Vague subjects are best avoided; many topics are best presented as questions:

> Should countries spend huge sums of money on space exploration?
> Do demonstrations serve any useful purpose?
> Do people ever really learn anything from the mistakes they make?

Group discussion and role playing

Group discussion and role playing are two other important techniques for assessing oral production. Through group discussion and role playing the teacher can discover how students are thinking and using the target language. For example, are they using the language they are learning to explore concepts and ideas? Or are they simply using the foreign language to present ideas already well-formed? In this way, group activities in both teaching and testing can be used to provide an opportunity for meaningful and active involvement. Students are thus given an opportunity to use what can be termed 'exploratory talk':[9] i.e. the language people use when trying to communicate rather than when they are engaged in the mechanical production of verbal formulae or patterns. In short, language becomes a means to an end rather than an end in itself. Group language activities then become ego-supporting (unlike the ego-threatening experience of too many students in traditional tests), as the other members of the group are interested chiefly in the message rather than in formal correctness.

Several useful books have been written on group work (dealing with the size and nature of various groups) and on activities for group discussion. Generally speaking, the type of activity most suitable for group

work is that in which the level of difficulty involved makes the task a little too difficult for the individual student to accomplish alone, but not so difficult as to discourage a group of students (especially given all the constraints and tensions of a test). In addition to all the problem-solving activities and puzzles which abound in books and materials of various kinds, the type of task involving concensus-seeking is particularly suitable for group discussion. In this latter type of activity, the members of the group are given a particular situation and instructed to make various decisions. It is necessary for them to use the target language to justify their decisions and seek agreement from the other members of the group.

> [10]Radioactivity from a nuclear power station accident will reach your area in a few hours. There is a small but very safe nuclear fallout shelter nearby, but there is room for only six people out of a total of twelve. Which six people from the following list do you think it would be most useful to save in the interests of future generations? List them in order of priority. (Note: M = male; F = female.)
>
> – a marine biologist, aged 56 (F)
> – a physicist, aged 25 (M)
> – a famous musician, aged 38 (F)
> – a farmer, aged 32 (M)
> – an electrician, aged 49 (M)
> – a mathematics teacher, aged 34 (F)
> – a well-known footballer, aged 22 (M)
> – a doctor, aged 63 (F)
> – a university student of sociology, aged 19 (F)
> – a fireman, aged 33 (M)
> – a factory worker, aged 28 (F)
> – a garage mechanic, aged 27 (M)

Role-play activities can also be used successfully to test oral communicative ability. The students involved are assigned fictitious roles and are required to improvise in language and behaviour. It is advisable for the students to be given fictitious names before the role play, as these usually prove very helpful in encouraging them to act out the roles assigned to them. The role plays used for the test may vary from short simple role plays involving only two or three students to far longer role plays involving several students. The students needn't be informed that they are being observed and assessed in their use of English: in fact, it is usually important that the teacher or examiner does this as discreetly as possible.

The following is an example of an extremely simple role play suitable for use at the elementary level.

> One student acts the part of a police officer, another a bus conductor, a third a bus-driver, a fourth a passenger hurrying to visit a sick friend in hospital, and a fifth a bystander who wants to be helpful. The passenger hurries to get on the bus and jumps on as it is moving off. The conductor stops the bus and tells him that the bus is full and that he must get off. The passenger can see an empty seat and he begins to argue. The bus is now in the middle of the road and is a danger to other traffic. Act the roles given.

It is usually advisable for each 'character' to be given a card on which there are a few sentences describing what kind of a person he or she is, etc Moreover, it is often helpful if the examiner can take a minor role in order

to be able to control or influence the role play if necessary. Indeed, even in group discussions and activities, the examiner should always feel free to interrupt or control the discussion in as diplomatic a way as possible in order to ensure that each member of the group makes a contribution. Both group discussions and role plays are best assessed if they are included as part of the language programme rather than as a formal test.

General conclusions

Generally speaking, a reliable method of obtaining measurements of oral production skills is that which involves the students' class teacher. The tensions and artificialities that inhibit the students in an oral examination conducted by an external examiner will now be avoided since the teacher is a familiar figure and the classroom a realistic part of the students' life. Continuous assessment by the teacher, with all his or her classroom experience, is generally (but not always) a reliable method of measuring the oral skills. Yet the oral interview (conducted by a sympathetic examiner) is still a useful examination to retain, particularly for its beneficial backwash effects on teaching. A comprehensive and balanced examination of oral production might thus consist of:

- an oral interview involving two students;
- a short problem-solving activity involving the comparison or sequencing of pictures, etc.;
- a longer activity comprising group discussion (concensus-seeking activity) or a role play.

The first two components listed here may be given as part of a formal test while the latter component is much better given during the course itself and assessed by the class teacher in as informal a way as possible. The examiner must frequently consider the effects of the examination on teaching and learning, however, and if, for instance, the reading aloud section is considered harmful in its effect on teaching, then it should be omitted from the examination.

Notes and references

1 This type of question has been used effectively in the *ARELS Oral Examination* (Association of Recognised English Language Schools).

2 The example is based on an item type in the University of Cambridge Local Examinations Syndicate: *Preliminary English Test* (revised version).

3 This technique is derived from a similar one used by A. S. Palmer and described in the article 'Testing Communication' in *IRAL* (*International Review of Applied Linguistics in Language Teaching*), Issue X/1, February 1972.

4 Richardson, R (ed.) 1976 *Learning for Change in World Society*. One World Trust

5 Based on the marking scheme devised by Dr Frank Chaplen for use in assessing the oral skills of overseas students at British universities. (The scheme is, in turn, derived from the American Foreign Service Institute rating scale.)

6 Morrow, K 1982 Testing Spoken Language. In *Language Testing* (ed. Heaton). Modern English Publications

7 Royal Society of Arts, Tests of Oral Interaction: Degrees of Skill. *The Communicative Use of English as a Foreign Language.*

8 Bartz and Schulz, 1974

9 Barnes, D 1973 *Language in the Classroom*. Penguin

10 Based on an idea in Richardson, R (ed.) 1976 *Learning for Change in World Society*. One World Trust

8

Testing reading comprehension

8.1 The nature of the reading skills

Until recently the many and diverse reading skills and strategies for use in everyday situations have been largely subordinate to a narrower range of skills required for dealing with simplified readers, especially at the elementary levels. Furthermore, on a few language courses, efficient reading skills have been pushed into the background in an attempt to develop oral fluency skills. Attempts at dealing with the many complex reading skills frequently come too late, at the tertiary level (i.e. at university, technical college), when students suddenly find themselves confronted with professional and technical literature in the foreign language.

In spite of the wide range of reading material specially written or adapted for English language learning purposes, there are few comprehensive systematic programmes which have been constructed from a detailed analysis of the skills required for efficient reading. Much test material is still limited to short reading extracts on which general 'comprehension' questions are based. As with listening comprehension, reading comprehension test material is very closely related to the type of practice material used by the teacher to develop the reading skills. Few language teachers would argue against the importance of reading: what is still urgently required in many classroom tests is a greater awareness of the actual processes involved in reading and the production of appropriate exercise and test materials to assist in the mastery of these processes.

Before reading tests in the second or foreign language can be successfully constructed, the first language reading skills of the students must be ascertained. Clearly there is often little purpose in testing in the second language those basic reading skills which the students have not yet developed in their own language. However, the mere fact that a student has mastered some of the required reading skills in the first language is no guarantee at all that he or she will be able to transfer those skills to reading another language.

At this stage in our examination of reading difficulties, it would be helpful to attempt to identify some of the specific skills involved in reading.[1] Broadly speaking, these can be defined as the ability to:

- recognise words and word groups, associating sounds with their corresponding graphic symbols;

– deduce the meaning of words by
  (a) understanding word formation (roots, affixation, derivation and compounding);
  (b) contextual clues (e.g. *One of the members of the group exposed the plot, and the police were soon able to arrest the leaders.*);
– understand explicitly stated information (e.g. *I wish Ann had come. = Ann did not come – hence my wish.*);
– understand relations within the sentence, especially
  (a) elements of sentence structure
  (b) negation
  (c) fronting and theme
  (d) complex embedding;
– understand relations between parts of a text through both lexical devices (e.g. repetition, synonyms, antithesis) and grammatical cohesive devices, especially anaphoric and cataphoric reference (e.g. *he, they, it; the former, the latter*) and connectives (e.g. *since, after, because, although, however, in addition*);
– perceive temporal and spatial relationships, and also sequences of ideas;
– understand conceptual meaning, especially
  (a) quantity and amount
  (b) definiteness and indefiniteness
  (c) comparison and degree
  (d) means and instrument
  (e) cause, result, purpose, reason, condition, addition, contrast, concession;
– anticipate and predict what will come next in the text;
– identify the main idea and other salient features in a text;
– generalise and draw conclusions;
– understand information not explicitly stated by
  (a) making inferences (i.e. reading between the lines)
  (b) understanding figurative language;
– skim and scan (looking for the general meaning and reading for specific information);
– read critically;
– adopt a flexible approach and vary reading strategies according to the type of material being read and the purpose for which it is being read.

No mention has been made here of reading aloud, since this particular skill is unique in that it involves different skills from silent reading.

Two different kinds of complementary reading activities to which students are usually exposed are generally classified as *intensive* and *extensive* reading. Short reading extracts of a moderate degree of difficulty and containing features which merit detailed study form a basis for intensive reading practice. Whole articles, chapters and books (usually simplified readers) are used for extensive reading practice; in this case, however, the material selected is generally slightly below the language attainment level of the students using it. Unfortunately, most reading tests concentrate on intensive reading to the exclusion of extensive reading, probably because it is more economical to have a large number of items based on a short reading extract than a few items based on a much longer one. However, these are insufficient grounds for neglecting to test extensive reading at certain levels.

In most tests, especially tests of general proficiency, it is useful to include a variety of text types for reading comprehension in addition to the

usual, more literary prose extracts: e.g. newspaper articles, instructions for using appliances and machinery, directory extracts, public notices, timetables and maps, advertisements, etc. The inclusion of such text types will not only provide a more realistic and reliable means of assessment but will also help to motivate students by demonstrating how the target language is used in real-life situations. Consequently, it becomes important that the actual presentation of the material should be as authentic as possible. In short, a newspaper article should appear in the actual form of a newspaper article, thereby giving a genuine feel to the material.

Several ways of testing reading comprehension are treated in the following sections of this chapter. Certain of the item types will be more suited to testing comprehension of a particular text than other item types. Indeed, there are numerous ways of testing reading comprehension, ranging from multiple-choice items to open-ended questions (i.e. questions which require students to write an answer in a sentence of their own). Although multiple-choice items are sometimes the most suitable instruments for testing reading comprehension, they should not be over-used. Frequently, other item types are far more interesting and useful. The text itself should always determine the types of questions which are constructed. Certain texts may lend themselves to multiple-choice items, others to true/false items, others to matching items, others to re-arrangement items, others to ordinary completion items, others to the completion of information in tables, and yet others to open-ended questions. Indeed, sometimes the same text will demand at least two or three different types of items.

8.2 Initial stages of reading: matching tests

The tests described in the first half of this section are concerned purely with word and sentence recognition. They test students' ability to discriminate visually between words which are spelt in fairly similar ways. If used in exercise material and progress tests, these test items will assist in developing word recognition speed. Though not administered as speed tests in the strict sense in the very early stages, word and sentence matching items should be covered by the students as quickly as possible. Once the students have gained familiarity and confidence with this type of test, their performances should be timed so that they are forced to read under some pressure. At first, it is advisable to confine the words used in the items to those already encountered orally; later a number of words not encountered orally should be introduced.

Word matching

The testees are required to draw a line under the word which is the same as the word on the left.

now	bow/not/how/<u>now</u>/mow
sheep	shop/shape/sleep/heap/sheep
ever	never/over/ever/fewer/even
top	top/stop/tap/pot/ton
wonder	wander/wonder/window/fonder/won
has gone	is gone/has won/has gone/his game/had gone
clothes	cloth/clothing/cloths/clots/clothes
most pleasant	most pleasant/more pleasant/most present/not pleasant/most peasants

Sentence matching

This item is similar to the word-matching item. The testees are required to recognise as quickly as possible sentences which consist of the same words

in the same order. They read a sentence, followed by four similar sentences, only one of which is exactly the same as the previous one.

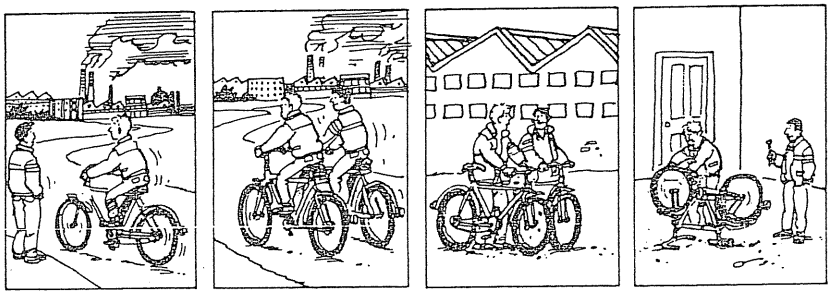
1. Tom is not going to your school.
   A. Tom is not going to your pool.
   B. Tom is going to your school.
   C. Tom is not coming to your school.
   D. Tom is not going to your school.

2. The thief can hide in the jungle.
   A. The thief can die in the jungle.
   B. The thieves can hide in the jungle.
   C. The thief can be hidden in the jungle.
   D. The thief can hide in the jungle.

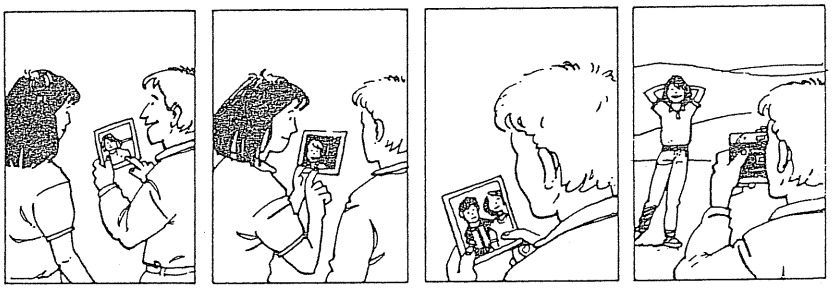
Pictures and sentence matching

In the remainder of this section the items will concentrate on word and sentence comprehension, using pictures to test this skill.

Type 1 This type of item is similar to that used to test listening comprehension and described under Type 3 in Section 6.5. The testees look at four pictures and then read a sentence about one of the pictures. They are required to identify the correct picture.

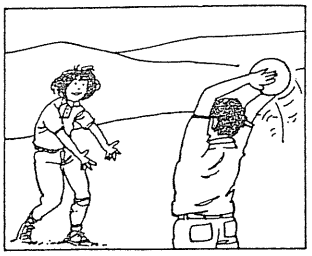


They are cycling to work.

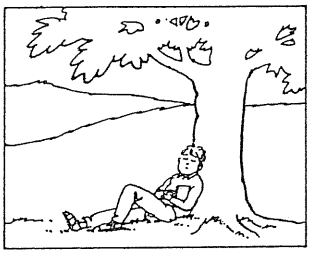


He is showing her the photograph.

Type 2 This type is similar to the previous one but is much more economical in that only one picture is required for each item (instead of four). The testees look at a picture and read four sentences, only one of which is about the picture. They then have to select the correct sentence.

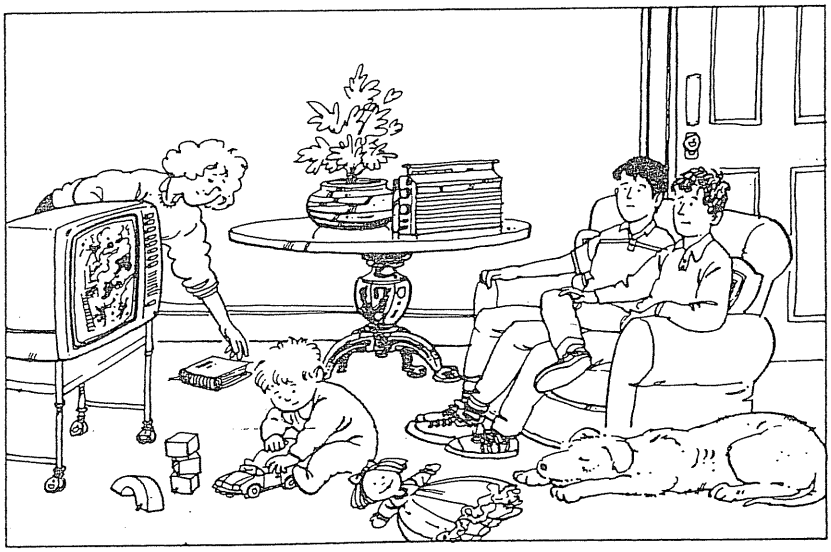


A. Jenny is throwing the ball to Peter.
B. Peter is kicking the ball to Jenny.
C. Peter is throwing the ball to Jenny.
D. Jenny is kicking the ball to Peter.



A. The man under the tree is reading his book.
B. The man resting under the tree is looking at his book.
C. The man with the book is sleeping under the tree.
D. The man carrying the book is going to sit down under the tree.

Type 3 Although this item type is referred to here as a matching item, it could equally well take the form of a true/false item (in which the testees write T or F at the side of each sentence according to whether or not the sentence agrees with the contents of the picture). In this particular instance, testees have to select the (four) sentences which match the picture.



Four of the following sentences agree with the picture. Put a circle round the letter of each of the four sentences.

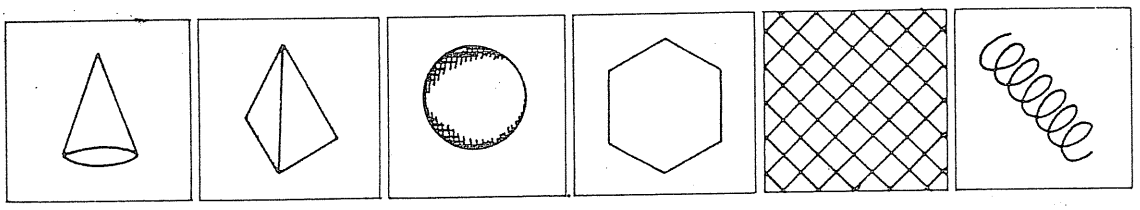
A. The dog on the floor is asleep.
B. The baby is playing with the dog.
C. The baby has just broken a toy car.
D. The television set is on fire.

E. The dog is behind the baby.
F. The woman has taken the flowers out of the bowl.
G. One of the two boys is helping the woman.
H. The woman is going to pick up a book.
I. The two boys are listening to the radio.
J. The radio is on the table but the book is on the floor under the table.

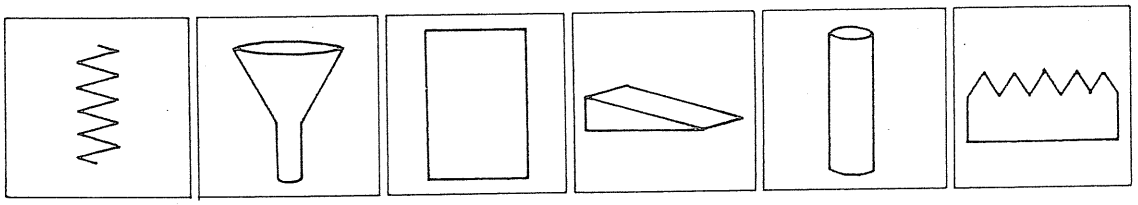
8.3 Intermediate and advanced stages of reading: matching tests

Type 1 The following matching item[2] shows how visuals can be used to test the comprehension of definitions of certain words. Testees are required to match the meaning of certain terms in a dictionary with the appropriate shapes which those terms denote. The example shows how the matching technique can be used at a more advanced level and how it can lend itself to a more communicative testing of reading. Above all, this particular item measures the ability of the testee to understand the kinds of definitions usually found in a dictionary – an essential skill required in learning and using a foreign language.

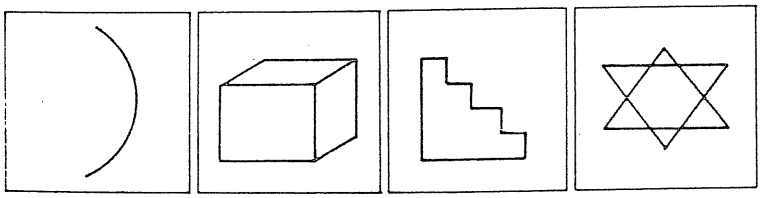
Some of the shapes are described in the dictionary extracts. Name *only* those that are described in the extracts. This first one has been done for you.



A Cone ..... B ..... C ..... D ..... E ..... F .....



G ..... H ..... I ..... J ..... K ..... L .....



M ..... N ..... O ..... P .....

coil /kɔɪl/ *vt,vi* wind or twist into a continuous circular or spiral shape; curl round and round: *The snake ~ed (itself) round the branch.* □ *n* [C] **1** something coiled; a single turn of something coiled: *the thick ~s of a python.* **2** length of wire wound in a spiral to conduct electric current.

cone /kəʊn/ *n* [C] **1** solid body which narrows to a point from a round, flat base. **2** something of this shape whether solid or hollow. **3** fruit of certain evergreen trees (fir, pine, cedar).

cube /kjub/ *n* [C] **1** solid body having six equal square sides; block of something so shaped or similarly shaped. **2** (*maths*) product of a number multiplied by itself twice: *The ~ of 5* (5^3) is 5 × 5 × 5 (*125*). □ *vt* multiply a number by itself twice: *10 ~d is 1 000.*

cu·bic /ˈkjubɪk/ *adj* having the shape of a cube; of a cube: *one ~ metre*, volume of a cube whose edge is one metre.

cu·bi·cal /ˈkjubɪkəl/ *adj* = cubic (the usual word).

cyl·in·der /ˈsɪlɪndə(r)/ *n* [C] **1** solid or hollow body shaped like a pole or log. **2** cylinder-shaped chamber (in an engine) in which gas or steam works a piston: *a six-~ engine/motor-car.*

cy·lin·dri·cal /sɪˈlɪndrɪkəl/ *adj* cylinder-shaped.

el·lipse /ɪˈlɪps/ *n* [C] regular oval.

el·lip·tic /ɪˈlɪptɪk/, **el·lip·ti·cal** /-kəl/ *adj* shaped like an ellipse.

funnel /ˈfʌnəl/ *n* [C] **1** tube or pipe wide at the top and narrowing at the bottom, for pouring liquids or powders through small openings. **2** outlet for smoke of a steamer, railway engine, etc. □ *vt, vi* (-ll-, *US* -l-) (cause to) move (as if) through a funnel.

lat·tice /ˈlætɪs/ *n* [C] framework of crossed laths or metal strips as a screen, fence or door, or for climbing plants to grow over: *a ~ window.*

lat·ticed *adj*

pyra·mid /ˈpɪrəmɪd/ *n* [C] **1** structure with a triangular or square base and sloping sides meeting at a point, esp one of those built of stone in ancient Egypt. **2** pile of objects in the shape of a pyramid.

sphere /sfɪə(r)/ *n* [C] **1** form of a globe; star; planet. ***music of the spheres,*** (*myth*) music produced by the movement of heavenly bodies which men cannot hear. **2** globe representing the earth or the night sky.

spheri·cal /ˈsferɪkəl/ *adj* shaped like a sphere.

wedge /wedʒ/ *n* [C] **1** V-shaped piece of wood or metal, used to split wood or rock (by being hammered), to widen an opening or to keep two things separate, ***the thin end of the wedge,*** (*fig*) a small change or demand likely to lead to big changes or demands. **2** something shaped like or used like a wedge: *~ heels (on shoes).* □ *vt* fix tightly (as) with a wedge: *~ a door open*, by placing a wedge under it. *be tightly ~d between two fat women on the bus.*

Type 2 The following item type[3] is included to provide another example of how reading comprehension matching tests can be based on the dictionary. Again, the item is intended for use at a fairly advanced level.

Settlements

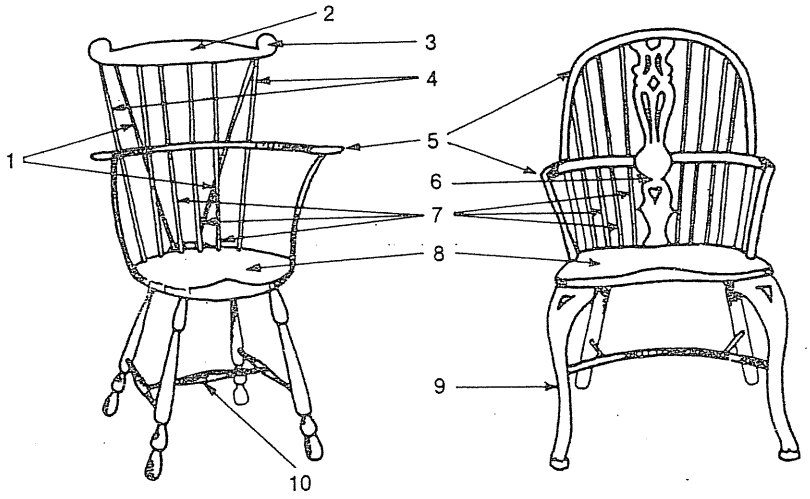
Enclosed hut groups are characteristic settlements in the area and remains of more than a hundred still exist. The open settlements, the villages of predominantly unenclosed huts, are not numerous but only a dozen groups have sufficient numbers of co-ordinated huts to be described as villages. Though there may be some walls in these villages, they are only fragmentary.

Many enclosed settlements have disappeared but one still extant is Rides Rys. It consists of two enclosures, one roughly square and a larger one roughly oblong in plan with a shared wall. An area of six acres was enclosed containing more than thirty buildings. The large enclosure, as in the case of other multiple settlements, had been built on to the smaller and indicates an expanding community.

The following dictionary definitions are for words used in the passage *Settlements*. Write the words from the passage next to the appropriate definition.

having or involving more than one part as an individual .....
serving as representative .....
becoming greater in size .....
enough to meet a need or purpose .....
having something in common .....
in most cases .....
is a sign of .....
very near to; approximate .....
broken off or incomplete .....
in existence; surviving .....

Type 3 The following item[4] is similar to the previous type but here testees are required to match appropriate words or information in the text with the correct parts of a diagram.



1 ..... 6 .....
2 ..... 7 .....
3 ..... 8 .....
4 ..... 9 .....
5 ..... 10 .....

Type 4 This item type involves the matching of a paraphrase of phrases and sentences in a text with the original words in the text, each item starting with 'Instead of . . .'. Although of use occasionally in achievement and proficiency tests, this type of item is more suited to teaching than to testing as it can be used to help students to increase their understanding of a text.

In the following example, students are asked to read a text and then complete each sentence. The item, however, calls for the matching of information and is thus not actually a completion type of item.

It is important for each student to obtain at regular intervals a rough idea of his or her progress. How are goals being achieved week by week? Methods of continuous assessment of students' work are replacing examinations – or parts of examinations – on certain courses. There are still doubts about the advantages of continuous assessment in the learning process but, if applied with care and discretion, continuous assessment can be a far more valuable means of assessing standards than an examination. Provided that methods of continuous assessment do not impart a feeling of tension and strain, they can be used to guide students in their work and to inform them of the progress they are making. If no means of continuous assessment is available, students should attempt to evaluate and summarise their progress very briefly week by week. Clearly, such an attempt is more difficult in a subject which teaches skills (e.g. learning a language, playing a musical instrument) than in a content subject (e.g. history, chemistry). Even as far as skills are concerned, however, it is a simple matter for students to go back to an old exercise and do it again. The ease with which they can do what previously seemed a difficult exercise is often quite remarkable.

1. Instead of talking about using such methods carefully and wisely, the writer talks about applying them .....
   .....
2. Instead of saying that it is harder for the students to assess progress made in learning a language than in history, the writer says that .....
   .....
3. Instead of referring to ways of measuring students' progress at regular intervals, the writer uses the phrase .....
   .....
4. Instead of saying that students should try to assess and report briefly on their progress every week, the writer says that .....
   .....
5. Instead of saying that continuous assessment is useful unless it makes students feel upset and worried, the writer says that it is useful
   .....

8.4 True/false reading tests

The true/false test is one of the most widely used tests of reading comprehension. Not only is the scoring of such a test straightforward and quick, but the scores obtained by the testees can be very reliable indices of reading comprehension provided that the items are well constructed and that there are enough of them. True/false tests are of considerable use for inclusion in class progress tests chiefly because, unlike multiple-choice test items, they can be constructed easily and quickly, allowing the teacher more time for other tasks.

The true/false test, however, has two main disadvantages: firstly, it can encourage guessing, since testees have a 50 per cent chance of giving a correct answer for each item. Secondly, as the base score is 50 per cent (= 0 per cent) and thus the average test difficulty generally in the region of 75 per cent (= 50 per cent), the test may fail to discriminate widely enough among the testees unless there are a lot of items.

It is, of course, possible to penalise the testee for guessing, and instructions on the lines of the following may be included in the rubric:

Each correct answer will be awarded two marks. However, for each wrong answer, one mark will be deducted from your score. It is better, therefore, not to guess blindly and to leave a blank if you do not know the correct answer.

Such penalties, however, are of dubious value, and the whole subject of guessing is treated in greater detail in Chapter 11.

Another solution to the problem of guessing is to include a third question in addition to the true/false option: e.g. true, false, not stated (i.e. the required information is not given in the passage). Thus, a rubric for this item type would read:

According to the passage, are the following statements true or false, or is it impossible to draw any conclusion?

In addition to the ease and speed with which the items can be constructed, the great merit of the true/false reading test lies in the ease with which suitable test passages can be selected: a short reading extract, for example, can provide a basis for numerous items. Moreover, the true/false test can be used as a valuable teaching device with which the students' attention is directed to the salient points in the extract by means of the true/false items.

If the students' comprehension of the true/false reading extract (and *not* the true/false items themselves) is being tested, each of the true/false items should be as clear and concise as possible. In such cases, it is essential that the problem posed by each item is fully understood. In many true/false reading tests some indication of the number of correct and incorrect statements is given to the students. Although this may make the test slightly easier for them, it does at least present them with a clear statement of the problem.

True/false reading tests fall into two general categories: those which are independent of a reading text (Type 1) and those which depend on a text (Type 2).

Type 1 It is possible to construct true/false items which are complete in themselves: a testee's comprehension of each true/false item is tested by means of a series of general truths. For example:

Put a circle round the letter T if the statement is true. If it is not true, put a circle round the letter F.

1. The sun rises in the west. T F
2. Fish can't fly, but birds can. T F
3. England is as large as Australia. T F
4. When ice melts, it turns into water. T F

Type 2 The construction of true/false items based on a reading extract forms one of the most widely used types of reading tests. It is often used at elementary levels of reading comprehension, but it can be used equally effectively at more advanced levels. The following example illustrates its use at a fairly advanced level:[5]

Eye-gazing and eye-avoidance have meanings and patterns of profound significance. Gazing at others' eyes generally signals a request for information and perhaps affection, but embarrassment can result from too long a mutual gaze. In fact, in intimate situations, there seems to be an equilibrium involving proximity, eye contact, intimacy of topic,

and smiling. If one component is changed, the others tend to change in the opposite direction.

But the extended gaze seems to have a function much deeper than that of maintaining a balance or ensuring a smooth flow of conversation. It signals, not surprisingly, an intensification of relationship, not necessarily along amorous lines. It may be a threat, or a challenge for dominance.

A definite pecking order of dominance and submission emerges from the very first eye contact of strangers. Curiously, when conversation is possible, it turns out that the one who looks away first tends to be dominant. The averted eye is a signal that its owner is about to take the floor. When conversation is not possible, however, the first to look away will be the submissive one.

Abnormal use of eye contact or aversion may well indicate an abnormal personality. Adult schizophrenics tend to use their eyes at all the wrong points in a conversation, and the bold liar can hold a steady gaze far longer than his truthful colleague when both are caught in the same misdemeanour.

According to the passage, six of the following statements are true and six are false. It is not possible to draw any conclusions from the information in the passage about the remaining three statements. Put T F or ? in each box, as appropriate.

1. Looking at someone else's eyes or looking away from them means a person is thinking very deeply. ☐
2. We generally look towards another person's eyes when we want information or a sign of affection from that person. ☐
3. If two people look too long at each other's eyes, they will usually become embarrassed. ☐
4. We are usually puzzled by someone who frequently looks into our eyes during a conversation. ☐
5. When engaged in very friendly conversation, a couple will probably look less at each other's eyes the more they smile and the closer they sit. ☐
6. It is easier to talk to someone in a friendly way if the person we are talking to does not gaze too long at the speaker. ☐
7. Looking for a long time at the other person's eyes is only a means of continuing a conversation smoothly. ☐
8. When two people gaze for a long time at each other's eyes, it is a sign that they are going to argue. ☐
9. An extended gaze can signal a threat or a bid for authority over the other person. ☐
10. When two strangers meet, they use their eyes to control or influence the other or to show their surrender to the other's authority. ☐
11. When it is possible to talk, the first person who looks away is the one who submits to the other person. ☐
12. A person shows that he or she wants to talk by looking towards the other person's eyes. ☐

13. People with poor eyesight generally stand or sit very close to the person they are addressing. ☐
14. Abnormal people usually turn their eyes away from the other person's eyes more often than normal people do. ☐
15. It is possible to tell the difference between a liar and an honest person by their eye-gazing patterns when both are trying to tell a lie. ☐

The reading text in the preceding example contained language at a higher difficulty level than that used in the statements which followed. It is possible (though not common practice), however, to construct a relatively simple text followed by more difficult statements: in such cases, the comprehension problems will be contained in the statements themselves rather than in the text.

8.5 Multiple-choice items (A): short texts

Type 1 It can be argued that the type of item in this section is in many ways a test of vocabulary rather than of reading comprehension. These particular items, however, have been included here because it is felt that a comprehension of the text is generally of at least as much importance as an understanding of the meaning of the words for selection. This, of course, is true of any vocabulary item presented in context: however, here the emphasis is more on the correct understanding of the context. The following three examples show the use of this item type at elementary, intermediate and advanced levels respectively.

1. The eyes are wonderful teachers – even musicians, who deal with sound, learn as much by (*doing, playing, watching, practising*) as by listening.
2. The housewife who could not afford to buy clothes would spend hours at her spinning wheel, spinning her wool into yarn – a job which took little skill but required a lot of (*ability, patience, talent, wisdom*) and was done by the fireside during the long winter evenings.
3. Two-thirds of the country's (*fuel, endeavour, industry, energy*) comes from imported oil, while the remaining one-third comes from coal. Moreover, soon the country will have its first nuclear power station.

Type 2 Just as the previous item type is closely related to the testing of vocabulary, so this type is perhaps more accurately described as a test of comprehension of grammatical structure. The testees are required to identify the correct paraphrase of a statement from a choice of four or five. They are told in the rubric that the (four) statements may refer to the entire sentence or only part of the sentence. Again, examples are provided for each of the three general levels.

1. John is not as tall as Sally but he's a little taller than Rick.
   A. Sally is taller than John and Rick.
   B. John is not as tall as Rick.
   C. Sally is taller than John but not as tall as Rick.
   D. Rick is taller than John and Sally.

2. In spite of the loud music, I soon managed to fall asleep.
   A. The loud music soon helped me to fall asleep.
   B. I soon fell asleep as a result of the loud music.
   C. The loud music made me unable to fall asleep soon.
   D. I soon fell asleep even though the music was loud.

3. If you'd forgotten to put out your hand, you wouldn't have passed your driving test.
   A. You didn't forget to put out your hand and you passed your driving test.
   B. You forgot to put out your hand and you failed your driving test.
   C. You forgot to put out your hand but you passed your driving test.
   D. You didn't forget to put out your hand but you didn't pass your driving test.

Type 3 This item type consists of a very short reading extract of only a few sentences (or sometimes of only one sentence). The testees are required to answer only one comprehension test item on each reading passage. The actual construction of multiple-choice reading comprehension items based on a reading extract will be treated in greater detail in the next section. Meanwhile, here are two examples of the use of multiple-choice items for testing reading comprehension, the first being at a fairly elementary level and the second at a more advanced level.

1. The president was talking to a young woman in the crowd when Tim suddenly caught sight of a man standing several yards behind her. The man had something in his hand: it was a short stick.

   What made Tim notice the man in the crowd?
   A. He was very close to Tim.
   B. The president was talking to him.
   C. He was standing in front of the woman.
   D. He was carrying a stick.

2. There were only two ways of moving along the narrow ledge: face outwards or face to the wall. I concluded that even the smallest of bottoms would push a person with his back to the wall far enough out to overbalance him and so, with arms outstretched in the shape of a cross and with chin pointed in the direction I was heading, I inched my way along.

   The writer managed to cross the narrow ledge by
   A. crawling along on his knees with his arms stretched out in front o him.
   B. moving sideways inch by inch with his back to the wall.
   C. working his way forward on his stomach with his face almost touching the ledge.
   D. walking slowly with his face and stomach close to the wall.

8.6 Multiple-choice items (B): longer texts

The multiple-choice test offers a useful way of testing reading comprehension. However, not all multiple-choice reading tests are necessarily good tests of reading comprehension. As was clearly indicated earlier, the extent to which a test is successful in measuring what it sets ou to measure depends largely on the effectiveness of each of the items used. Indeed, certain general aspects of many reading tests may be suspect. For instance, does the usual brief extract for reading comprehension

concentrate too much on developing only those skills required for intensive reading, encouraging frequent regressions and a word-by-word approach to reading?

The sampling of the reading passage is of the utmost importance and must be related to the broader aims of the language teaching situation. Many of the texts in both school and public examinations concentrate too much on a literary kind of English. If certain students are learning English in order to read technical journals, for example, then the sampling of the reading extract should reflect this aim. Ideally, in a test of proficiency the text should contain the type of reading task which will be demanded of the testees in later real-life situations. If the test is a class progress or achievement test, the reading passage should be similar to the type of reading material with which the students have been confronted in their work at school. In other words, if other subjects are being taught in the medium of English (as in many second language situations), the text should frequently (though not *always*) reflect the type of reading the students are required to do in history or chemistry, etc.

In this section, it is assumed that only intensive reading skills are being tested. Thus, the length of the reading extract recommended might vary from 50 to 100 words at the elementary level, 200 to 300 words at the intermediate level, and 400 to 600 words at the advanced level. These figures are, of course, extremely rough guides and may not be appropriate for many reading situations. Moreover, the extract selected should be capable of providing the basis for a sufficient number of multiple-choice comprehension items. It is not an easy task to find an extract which will support a number of multiple-choice items – even though the same extract may form a basis for a large number of true/false items or open-ended questions. Generally speaking, passages dealing with a series of events, a collection of facts, or different opinions and attitudes make the best types of texts for testing purposes; those dealing with a single idea or main theme are rarely suitable.

The length of the extract should also be related to its level of difficulty: a particularly difficult or complex passage would probably be considerably shorter than a more straightforward one. On the whole, the difficulty level of the text, however, should coincide with the level of the students' proficiency in English, but we must remember that the reading matter used outside the test situation (e.g. simplified readers) should be selected for enjoyment and should thus be at a slightly lower level than the actual standard of the reading skills acquired. (The difficulty level of a text depends chiefly on the degree of the structural and lexical complexity of the language used.)

When writing test items based on a reading text, the tester should attempt to construct more items than the number actually required. After the construction of the items, it is useful to secure the services of one or two colleagues so that all the items can be moderated. Invariably this process brings to the attention of the item writer certain flaws in some of the items. Although a number of the flaws will be easily rectified, in certain cases it will be necessary to dispense with entire items. In tests of grammar and vocabulary, new items can always be constructed in place of the discarded items, but this does not follow with reading comprehension items. The text itself has to be rewritten, certain sections added and others deleted in order to obtain the required number of items. Such processes are difficult and time-consuming: thus, it is always an advantage to

construct in the first instance more items than are actually required. If the text will not allow for more items, another more suitable text should be chosen to avoid wasting time at a later stage.

It is useful to include items testing the students' ability to recognise reference features in a text, no matter whether multiple-choice, completion or open-ended items are being constructed. A reference-word item can provide the examiner with specific information about reading difficulties. If a student fails to perceive what the reference device 'it' refers to in the text, for example, the examiner immediately knows the reason for his or her failure to understand that part of the text.

> The grizzly bear roams some 12 million acres in rugged parts of the United States. And this great bear still roams our imagination at will; *it* is part of its natural habitat.
>
> The word *it* in line 3 refers to
> A. 'the United States' (lines 1–2)
> B. 'this great bear' (line 2)
> C. 'our imagination' (line 2)
> D. 'its natural habitat' (line 3)

How many multiple-choice items should be set on one text? Clearly, the number of items will depend on the length and complexity of the text. However, tests of reading comprehension generally contain fewer items than other skill tests. Furthermore, the testees require much more time to work through a reading comprehension test since they first have to read the text carefully once or twice before they can begin to answer the items based on it. While as little as ten or fifteen seconds for each item can be allowed in multiple-choice tests of grammar and vocabulary, at least one or two minutes must be allowed for each item in the average reading test (if the time required to read the text is taken into account). Consequently, such tests, though long in terms of time, must of necessity be short in terms of items and, therefore, less reliable.

The construction of items depending simply on a matching of words and phrases should be avoided. Items should test more than a superficial understanding of the text and should require the testees to digest and interpret what they have read. The following examples show how ineffective items can be if testees are simply required to match the words in the items with the words in the text.

> At four o'clock on September 30th two men armed with iron bars attacked a soldier in Priory Street.
>
> What happened at four o'clock on September 30th?
> A. Two neminsi deraden with rinot babblers tacklened a derisoldt.

Imagine that a testee did not understand much of the sentence in the text. In order to appreciate this fully, it is necessary to change the situation slightly, and the text might appear to us like this:

> At four o'clock on September 30th two neminsi deraden with rinot babblers tacklened a derisoldt.
>
> What happened at four o'clock on September 30th?
> A. Two neminsi deraden with rinot babblers tacklened a derisoldt.

A slightly better item stem would be:

> What happened one afternoon at the end of September?

However, to be completely satisfactory, it would be necessary to rewrite both the text and the item, as in the following example:

> Paul was surprised when he met Sue at the party. He was under the impression she had gone away from the locality. The last time he saw her was when Jane was teaching her to drive. A few days afterwards she had suddenly become ill.

(first version)

Paul was surprised when
A. Sue went away.
B. he met Sue at the party.
C. Jane was teaching Sue to drive.
D. Sue suddenly became ill.

(second version)

Paul did not expect to see Sue because
A. he knew she was at the party.
B. he thought she had left the district.
C. he had seen Jane teaching her to drive.
D. he had heard she was ill.

There is often a temptation to concentrate too much on facts, figures and dates when constructing test items based on a factual text. Generally speaking, figures and dates are included in a text chiefly for the purpose of illustration or to show the application of a general principle. It is useful in such cases to construct items which require the testees to use the figures in the text to state (or restate) the general principle behind them. E.g.:

> From January to December last year, 291 people were killed and 6,248 were injured in road accidents on the city's roads. 157 of all the fatal accidents involved motorcyclists or their pillion passengers, while 95 involved pedestrians and the remaining 39 the drivers and passengers of motor vehicles.
>
> Over half of all the people killed in road accidents last year were
> A. motorcyclists and pillion passengers.
> B. pedestrians.
> C. drivers of buses, cars and lorries.
> D. both pedestrians and motorists.

Testees can also be encouraged to use the figures they are given in a text and to work out simple arithmetical sums and problems. Clearly, there is a limit to the tasks which the testees may be required to perform: otherwise the test writer will be testing something other than language skills. The following is an example of an item which tests students' ability to handle simple facts and figures in English: the stem presents a useful task *provided that this kind of reading exercise is not overdone.*

> Latest reports from the northeast provinces state that at least sixteen people lost their lives in Saturday's floods. A further nine people, mostly children, are reported missing, believed dead. Seven small boys, however, had a miraculous escape when they were swept onto the branches of some tall trees.
>
> The total number of people reported dead or missing as a result of Saturday's floods is
> A. 7 B. 9 C. 16 D. 25 E. 32

The choice of the correct option in each multiple-choice item must depend on a testee's comprehension of the reading text rather than on general knowledge or intelligence. The following item, for example, can be answered without any knowledge of the text on which it has been based.

Memorising is easier when the material to be learnt is
A. in a foreign language.
B. already partly known.
C. unfamiliar and not too easy.
D. of no special interest.

Care must be taken to avoid setting distractors which may be true, even though they may not have been explicitly stated by the writer. In the following test item based on a reading text about the United Nations and the dangers of war, C is the required answer; however, all four options are correct – even though not stated in so many words by the writer.

What would happen if there was a global war?
A. Nations would train men for war.
B. Lots of terrible weapons would be made.
C. The whole human race would be completely destroyed.
D. People would grow very desperate.

The correct option must be roughly the same length as the distractors. In the following test item the correct option has been modified to such a degree that it appears as the obvious answer without even necessitating any reference to the text.

The curriculum at the new college is a good one in many ways because it
A. includes many science courses.
B. offers a well-balanced programme in both the humanities and the sciences.
C. is realistic.
D. consists of useful technical subjects.

All the options must be grammatically correct: there is a tendency especially in reading comprehension to overlook the grammatical appropriateness of some of the distractors used. Option D in the following item can be ruled out immediately because it is ungrammatical.

The writer says that he had studied engineering for
A. a long time.
B. only a very short period.
C. several years.
D. never.

Double negatives are only confusing and such items as the following (based on the extract on page 120) are best avoided:

Paul did not expect to see Sue because
A. he did not know she was at the party.
B. no one knew she had left the district.
C. he hadn't seen Jane teaching her to drive.
D. he didn't realise she was well.

A useful device in multiple-choice tests of reading comprehension is the option ALL OF THESE or NONE OF THESE:

According to the passage, what do some people think there should be outside a modern city?

A. Buses
B. Car parks
C. Office buildings
D. Taxis
E. ALL OF THESE

If an option like E is used, it is advisable to have it as the correct answer in at least one of the items. The testees should not be encouraged to think that it has been included simply to make up the required number of options.

The following text and comprehension items[6] illustrate some of the guidelines laid down in this section:

Study the following passage and then answer the questions set on it.

The Captive is a strange but sincere and tender film, as indeed one would expect from a director of the calibre of Marcel Lyme. In addition to his keen sensitivity, Lyme has a strong feeling for historical atmosphere, so apparent in his earlier film *Under the Shadow of the Guillotine*, in which the events of the French Revolution are depicted with surprising realism and vitality. In *The Captive* Lyme manages to evoke the atmosphere of an English town in the early part of the nineteenth century, not so much through the more obvious devices of stage-coaches, old inns, and thatched cottages as through minute attention to details of speech, dress, customs, and mannerisms. Similar in theme to *Adam Brown*, *The Captive* is distinguished by a sincerity which the former lacks and which helps to transform this film from an ordinary adventure story into a memorable and a very moving tragedy. Especially unforgettable is the farewell scene at Plymouth, when Jonathan Robson sees Catherine Winsome on his way to the grim, squalid ship which is waiting to take him to Australia. Robson breaks loose from his captors for a fleeting moment to bid farewell to Catherine. 'I'll prove my innocence,' he cries vehemently as he shakes his fist at Catherine's cousin.

As the ship sets sail, one enters a grotesque nightmare world in which evil seems triumphant. Our identification with Robson becomes so personal that we feel every stroke of the flogging after he has been caught stealing medicine for his sick companion. We share his sympathy for Joe Biggs as the old sailor is hauled under the ship's keel. Indeed, events might well have become unbearable but for the light relief provided by the comical antics of Bobo, the small cabin boy who skips about uncomplainingly doing whatever task he is given. We know, of course, that ultimately evil will be vanquished, and so we are given strength to endure the adversities which confront the hero. The mutiny and the consequent escape of Jonathan Robson, therefore, come as no surprise.

Questions

(a) For each of the following statements choose the word or phrase that best completes the statement according to the information contained in the passage. Write the number of the question and the answer you have chosen in your answer book.

(i) *The Captive* was directed by
A. Jonathan Brown.
B. Adam Brown.
C. Marcel Lyme.
D. Catherine Winsome.

(ii) In *The Captive* Marcel Lyme conveys the atmosphere of the nineteenth century chiefly through
A. close attention to small details.
B. the use of conventional scenery.
C. stage-coaches, old inns, and thatched cottages.
D. depicting dramatic events of the time.

(iii) The passage implies that *Adam Brown* was
A. a very moving film.
B. a realistic and vital film.
C. an ordinary adventure film.
D. a sincere film.

(iv) Jonathan Robson is angry as a result of
A. having to wait to go to Australia.
B. being wrongly convicted.
C. meeting Catherine.
D. being recaptured.

(v) On the voyage to Australia Robson
A. becomes ill.
B. begins to have nightmares.
C. is hauled under the ship's keel.
D. receives a flogging.

(vi) Bobo is introduced into the story to help us to bear the grim events by
A. behaving in a strange but interesting way.
B. making us laugh.
C. doing everything without complaining.
D. acting kindly toward the hero.

(vii) We can endure the hero's sufferings because we know
A. things cannot get worse.
B. the crew will mutiny.
C. good will win in the end.
D. the hero is very brave.

(viii) The writer's attitude to this film is
A. appreciative
B. patronising.
C. scornful.
D. critical.

(ix) The word 'his' in line 3 refers to
A. '*The Captive*' (line 1)
B. 'one' (line 1)
C. 'a director' (line 2)
D. 'Lyme' (line 3)

(x) The words 'the former' in line 12 refer to
A. 'theme' (line 11)
B. '*Adam Brown*' (line 11)

C. *'The Captive'* (line 11)
D. 'a sincerity' (line 11)

(xi) The word 'his' in the phrase 'We share his sympathy' in line 23 refers to
A. 'Robson' (line 21)
B. 'his sick companion' (line 23)
C. 'Joe Biggs' (line 24)
D. 'the old sailor' (line 24)

(xii) The word 'he' in line 27 refers to
A. 'Robson' (line 21)
B. 'Joe Biggs' (line 24)
C. 'the comical antics' (line 26)
D. 'Bobo' (line 26)

(b) Each of the following words and phrases can be used to replace one word in the passage. Find the words and write them in your answer book. Number your answers.

(i) dragged
(ii) conquered
(iii) troubles and misfortunes
(iv) very brief
(v) finally

8.7 Completion items

Completion items measure recall rather than recognition. Although such items are similar in many ways to open-ended questions in tests of reading comprehension, they are often regarded as belonging more to the objective category of test items. There is very little difference, however, between the following open-ended reading comprehension question:

Why was the author surprised to meet Dr Short?

and the equivalent completion item:

The author was surprised to meet Dr Short because .....

Usually, completion items require the testees to supply a word or a short phrase. Unless great care is taken to ensure that there is only one correct answer, the marking will prove very difficult when the tester is confronted with a variety of answers ranging from acceptable to unacceptable. All valid interpretations, whether or not these were in the test writer's mind at the time of the construction of the test, must be regarded as correct.

Types of completion items for testing reading comprehension are divided into two groups for ease of treatment: Type 1 consisting of blanks for completion in the items following the text, and Type 2 consisting of blanks in the text itself.

Type 1 Unless carefully constructed, this type of completion test can become merely a matching exercise in which the words and phrases required in the completion are determined after a process of matching the whole item with the appropriate part of the text. However, less emphasis is placed on the writing skills in such a test than in a test consisting of open-ended questions. The first example of this item type illustrates how a short informal letter can be used to provide the basis for completion items.

Example 1

256 Weeton Road,
2nd Floor,
Hong Kong.
7th June, 1988.

Dear David,
I am very sorry that I could not meet you last night. I hope that you did not wait too long outside the New York Theatre. I had to look after my small brother until my mother returned home. She was a long time at the doctor's and she arrived home very late. I ran all the way to the bus stop, but I had already missed the bus. I decided to get on a tram and I arrived at the New York Theatre at eight o'clock. I did not think that you would still be there because I was three-quarters of an hour late. I do hope that you will forgive me.

Your friend,
Peter

Write one word or more in each blank.

1. Peter lives at .....
2. He wrote the letter on .....
3. Peter could not leave home because he had to wait for ..... to return.
4. His mother had been to the .....
5. Peter went to the New York Theatre by .....
6. He thought that David .....
7. The word ..... means *excuse*.
8. Peter had arranged to meet David at ..... seven on June .....

Example 2 The second example gives the stimulus in the form of a theatre advertisement. It has been included here as a reminder of the importance of varying text types in a reading comprehension test and of using the type of reading material which the student may meet in places where English is used as an everyday means of communication.

Read the following advertisement and complete each sentence. Write one word or more in each space.

NEWTON THEATRE
FORTHCOMING ATTRACTIONS

Monday, 8 January for 2 weeks
MY FAT FRIEND
Charles Laurence's popular comedy

Wednesday, 24 January to Saturday, 27 January
Shanghai Festival Ballet
presents
SWAN LAKE

Monday, 29 January for one week
RUN AND KILL
Tim Danby's thrilling mystery

1. The Shanghai Festival Ballet will perform on ..... evenings.
2. ..... will be the most amusing play.
3. If you like dancing, you should see .....
4. The play written by ..... is very exciting.

Example 3 In this completion item the students are required to fill in a table based on a passage for reading comprehension. The information extracted from the text is best tested by requiring the students to put it in tabular form.

> The city with the highest temperature yesterday was Singapore. At noon the temperature in Singapore was 33°C and at midnight the temperature there was 25°C. Tokyo had the second highest temperature. It was only 2°C lower there than in Singapore at noon. The temperature in Rome at noon was 30°C, the same as in New York. However, New York's temperature at midnight was one degree lower than Rome's. The noon temperature in Cairo was 29°C, one degree higher than that in Athens and Hong Kong. The temperature at midnight in Paris was 11°C less than that at noon in Paris yesterday. Although Stockholm's temperature at midnight was the same as the temperature at midnight in Paris, its noon temperature was 14°C. The coolest city was London, with a temperature of only 22°C at noon and 13°C at midnight.
>
> The five cities with the highest temperatures had a lot of sunshine throughout the day. Although the sun did not shine at all in Athens and Paris, it did not rain. It was also cloudy and dry in Stockholm, but it rained heavily in both London and Hong Kong.

Yesterday's Weather Around the World

S = Sunny C = Cloudy R = Rain			
	Noon	Midnight	
		26°C	
Cairo		26°C	S
Hong Kong		25°C	
	22°C		R
	30°C		
Paris	25°C		
Rome		25°C	S
Singapore	33°C	25°C	S
Stockholm			C
Tokyo		26°C	S

Type 2 In this item type the testees are required to complete the blank spaces in a reading text. The blanks have been substituted for what the test writer considers are the most significant content words. Consequently, a possible weakness of such a test may result from the failure to supply adequate guidelines to the testees; the following is an example of a poor item because the framework is insufficient to guide them. The linguistic clues are thus inadequate and the testees are faced with the task of having to guess what was in the examiner's mind.

> When we (1) ..... something along the (2) ....., it will cause (3) .....

The following text illustrates how blanks should be interspersed; the testee's degree of success in completing the blanks depends almost entirely on his or her comprehension of the whole text.

> When we slide something along the floor, it will cause (1) ..... If something is very (2) ....., there will be a lot of friction between it and the ground. However, friction is (3) ..... when something rolls instead of (4) ..... The invention of the (5) ..... was really an attempt to reduce friction. Unless there is snow or ice, it is much harder to (6) ..... something on a box or sledge than in a cart. Ball-bearings are used a lot in machinery to (7) ..... friction. It is friction which causes (8) ..... to machinery as the various parts (9) ..... against one another. Friction is reduced if we put oil onto the (10) ..... It should not be forgotten, however, that (11) ..... is also useful to us: it is necessary, for instance, for car wheels to grip the (12) .....

Example 1 In some tests certain letters of missing words are given. In these cases, the testees are generally informed that each dash in the blank signifies a letter.

> The mighty Amazon f---s into the Atlantic near the Equator: its es----y is about 170 miles wide. The w--th-- is often so misty that the b---s of the river cannot be seen from a ship, even if it is p-s---g quite close to them.

It is also possible to provide only the initial letter of the missing word. This item is not to be recommended for most purposes as the inclusion of letters can often create mental blocks and only confuse students if they fail to think of the exact word required. This is clearly disadvantageous for reliable assessment as a student might understand the passage and would otherwise have been able to complete the passage with a suitable synonym.

Example 2 A variation of this type of reading comprehension may incorporate the multiple-choice technique:

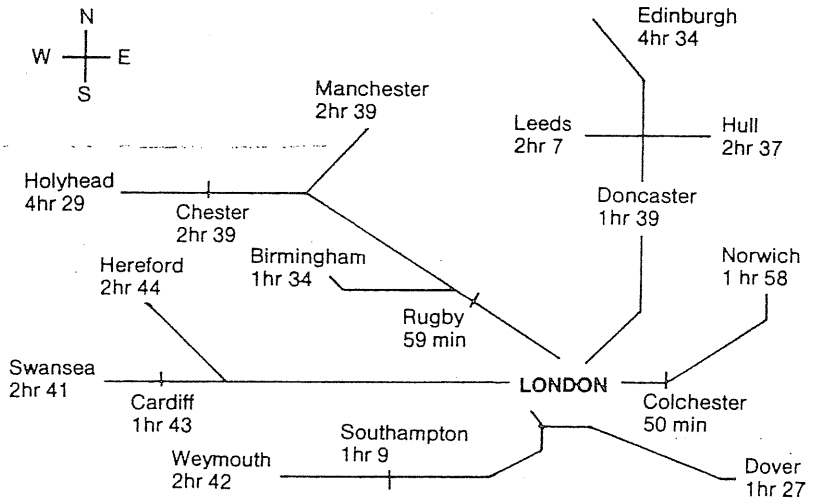
> Astrology is the ancient (1) ..... of telling what will (2) ..... in the future by studying the (3) ..... of the stars and planets. (4) ..... astrologers thought that the stars and planets influenced the (5) ..... of men, they claimed they could tell (6) .....
>
> (1) custom business magic knowledge art
>
> (2) coincide happen chance come foretell
>
> (3) places shapes times positions light
>
> (4) However Because Although For While
>
> (5) affairs matters businesses chances times
>
> (6) horoscopes future advice fortunes luck

Example 3 The following type of item is used to most advantage when the item itself is related to the kinds of tasks the testees are required to perform in their studies or in real life. For example, testees can be provided with the dictionary definitions of a number of words. (This material is simply extracted from a good dictionary, together with information concerning pronunciation, verb patterns, parts of speech, etc.

and examples of use where appropriate.) Underneath the dictionary extract are printed incomplete sentences, which testees are required to complete with the most appropriate word.

Example 4 In the following type of item the reading comprehension text is related to information contained in a diagram or table.

Look at the following diagram. It shows the places to which trains travel from London. It also shows the times taken to travel to these places.



It takes two hours and seven minutes to travel to (1) from London, half an hour less than it takes to (2). It takes one hour and thirty-nine minutes to travel to (3), which is also in the north of England. It takes the same time to reach (4) as it does to reach Manchester. (5) is the furthest place from London. (6), a small town in the west of the country, is only five minutes nearer London than Edinburgh is. Just under two hours' travelling time from London is (7), a city to the east of London. It takes a little under an hour and a half to reach (8), a pleasant town to the southeast of London. To the southwest of London is (9), which is one hour and nine minutes from there and is on the way to (10). If you travel to (11), it will take two hours and forty-one minutes. You will pass through (12) almost an hour before you reach there. (13) is the second largest city in England, one hour and thirty-four minutes from London. If you live in (14), you can catch the Birmingham train, the Manchester train or the Holyhead train. All three trains will pass through this lovely country town.

(1) .....
(2) .....
(3) .....
(4) .....
(5) .....
(6) .....
(7) .....
(8) .....
(9) .....
(10) .....
(11) .....
(12) .....
(13) .....
(14) .....

Example 5 Completion items are also useful for testing the ability to deduce word meanings from the clues available in the context. Such an ability is of the utmost importance in reading and should be encouraged in both teaching and testing at all levels. The following is an example of an

item designed to test this ability at the intermediate level. Students are instructed to read the entire text before attempting to complete each blank. The first example contains only one blank while the second example contains three blanks, each requiring the same word.

(One blank: one word)
The Great Pyramid at Giza in Egypt is without doubt still one of the greatest . . . . . in the world, even though it has been studied very thoroughly over the last century or so. It may, however, be possible to find out more about the pyramid by closer study over the next few years.

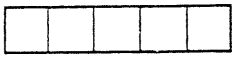
(Three blanks: one word)
Japan, in the interests of both the Japanese and the world . . . . . , needs to be assured that efforts will be made to stabilise the yen. For its part, Britain is now taking active steps to improve its . . . . . by developing new industries and cutting down on imported goods of all kinds. Talks now being held in Paris between France and Germany are directed at ways of increasing the manufacturing capabilities of the two countries in the hope that this will lead to a general improvement in the European . . . . . and create more jobs.

8.8 Rearrangement items

These two item types are particularly useful for testing the ability to understand a sequence of steps in a process or events in a narrative. While in an exercise for classroom practice the students will often be required to rewrite the jumbled sentences in their correct sequence, it is obviously preferable for testing purposes to instruct them to write simply the numbers or letters of the jumbled sentences. It is also advisable to provide them with one or two answers: if students start off by putting the first two or three sentences in the wrong order, it may be impossible for them to put the remaining sentences in the correct order. In other words, one wrong answer will inevitably lead to a second wrong answer and possibly a third, and so on. This weakness must be borne in mind when marking: for instance, should two incorrect answers (i.e. one incorrect answer automatically resulting in another) be penalised by the loss of two marks?

Type 1 The first of these items takes the form of jumbled sentences which the students have to 'unscramble' and arrange in the correct order. The students are instructed to rearrange the letters of the sentences in their correct order in the boxes below.

A. The dogs were separated from their human masters and were put into large fields.
B. For instance, they wagged their tails, they barked and growled, and they fawned on animals which possessed food.
C. Psychologists observing them found that they reacted towards each other in much the same way as they used to respond to people.
D. Recently an experiment was conducted with a group of dogs to find out how much their behaviour was, in fact, simply a result of human environment.
E. Puppies born to those dogs and kept out of contact with people showed the same behaviour patterns: they were extremely wiid and afraid of human beings.



Type 2 The jumbled sentences in the second item are based on a reading comprehension text and should be unscrambled in the light of the information contained in the text. The students must write the letters of the sentences in their correct order in the boxes at the end of the item. Again, however, the fact that one mistake inevitably results in another reduces the item's reliability.

When a customer walks into a travel agency to make a booking, the clerk behind the counter turns immediately to the small computer unit on the desk. The unit looks like a synthesis of a television set and a typewriter.

The customer states the date and destination of the flight required, and the clerk types this information on the keyboard. As each key is pressed, a letter is formed on the screen of the unit. The clerk then checks the information on the screen before transmitting the data to a central computer. This computer contains all the information about current bookings and destinations, and rapidly establishes whether the new booking is possible. The computer immediately sends a reply, indicating the number of vacancies on the flight requested or showing that the flight is fully booked. While these figures are being displayed on the screen, they may change to indicate that another booking has been made in another part of the world.

The clerk now types in the customer's reservation, after which the computer will request his or her name and address and then other information. This information, including an indication of how the ticket will be purchased (cash or credit card), is then typed onto the screen. Next the computer confirms the booking and requests that payment be made. When the customer has paid for the ticket, the clerk types this information into the computer as well. Finally, if modern equipment is being used, it is possible for the computer to print out a ticket on the spot.

Now put these sentences in the correct order. Write only the *letter* of each sentence in each box. (Three boxes have been completed for you.)

A. Details about the seats available are sent back.
B. The computer then wants to be informed about the method of payment.
C. The tickets are issued.
D. The computer asks for personal details.
E. The customer makes his or her request to the travel agent.
F. The request is sent to the main computer.
G. The customer goes to a travel agent.
H. The computer requests payment.
I. The travel agent feeds the initial request into a small computer unit.
J. The booking is typed into the computer.
K. The booking is confirmed.

G			F		J					

As in the case of true/false type reading items, it is possible to improve the reliability of the item by introducing an additional feature: i.e. 'Don't know', 'Information not given', 'Not applicable'. The following item[7] is based on an article contained in a separate pull-out sheet using a newspaper format (but not shown here).

You would like to learn something about migraine and so read the article 'One-sided Headache'.

In what order does the writer do the following in her article? To answer this, put the number 1 in the first answer column next to the one that appears first, and so on. If an idea does not appear in the article, write N/A (not applicable) in the answer column.

Write in every box

a. She gives some of the history of migraine.	a.
b. She recommends specific drugs.	b.
c. She recommends a herbal cure.	c.
d. She describes migraine attacks.	d.
e. She gives general advice to migraine sufferers.	e.

8.9 Cloze procedure

Although similar in appearance to completion items, cloze tests should not be confused with simple blank-filling tests. In ordinary completion tests the words for deletion are selected subjectively (consisting largely of structural words in tests of grammar and key content words in vocabulary or reading tests). In cloze tests, however, the words are deleted systematically. Thus, once the actual text has been chosen, the construction of a cloze test is purely mechanical: every *n*th word is deleted by the test writer. Certain test writers argue that the blank substituted for the deleted word should correspond to the length of the missing word but in most cloze tests all the blanks are now of uniform length. Unless a photostat copy of the actual printed text is being used (in which case the words are deleted before the photocopying process), it is simpler to insert blanks of uniform length.

The interval at which words are deleted is usually between every fifth and every tenth word. However, if every seventh word has been deleted in the first few sentences, then every seventh word must be deleted for the rest of the text. The fifth, sixth and seventh words are the most widely favoured for deletion in cloze tests. If every third or fourth word is deleted, the student will have extreme difficulty in understanding the text as insufficient clues will be available. If every tenth or twelfth word is deleted, it will be necessary to have a long text. For example, if there are 40 deletions every seventh word, the text will be 280–300 words in length (i.e. 40 × 7); if, however, there are 40 deletions every twelfth word, the length of the text will be 480–500 words (i.e. 40 × 12).

The cloze test was originally intended to measure the reading difficulty level of a text. Used in this way, it is a reliable means[8] of determining whether or not certain texts are at an appropriate level for particular groups of students (both native speakers and non-native speakers). The reading text being evaluated is given to a group of students and the average score of all the students in the group is obtained. If the mean score of the group is over 53 per cent, the material can be used by the students for reading at 'the independent level', the text being considered easy enough for students to read on their own without any help. If the mean score obtained is between 44 and 53 per cent, however, the material is suitable for use at 'the instructional level' – i.e. with the help of the

teacher. If the mean score is below 44 per cent, the text is described as being at 'the frustrational level' and should not be used even with the help of a teacher. Later research, however, has shown that this range of scores may vary considerably, the instructional level on some occasions ranging from 38 to 50 per cent and on other occasions from 47 to 61 per cent.

Perhaps the most common purpose of the cloze test, however, is to measure reading comprehension. It has long been argued that cloze measures textual knowledge: i.e. an awareness of cohesion in a text, involving the interdependence of phrases, sentences and paragraphs within the text.[8] Unlike the other types of items treated here, however, a true cloze test (i.e. the mechanical deletion of words in the text) is generally used to measure *global* reading comprehension – although insights can undoubtedly be gained into particular reading difficulties. However, it is also argued that cloze measures an underlying global linguistic ability (and indeed even knowledge of the world) rather than simply those skills associated with reading comprehension. As the cloze procedure is now such an important feature in language testing (especially in the United States), the whole subject is treated at much greater length in the following chapter.

Meanwhile, the following examples have been included in this section to demonstrate how cloze procedure can be applied to the testing of reading comprehension at both the elementary and the more advanced levels.

Example 1 (Elementary)

Once upon a time a farmer had three sons. The farmer was rich and had many fields, but his sons were lazy. When the farmer was dying, he called his three sons to him. 'I have left you . . . . . which will make you . . . . . ,' he told them. 'But . . . . . must dig in all . . . . . fields to find the . . . . . where the treasure is . . . . . .'

After the old man . . . . . ., his three lazy sons . . . . . out into the fields . . . . . began to dig. 'I'll . . . . . the first to find . . . . . place where the treasure . . . . . buried,' cried the eldest . . . . . 'That's the field where . . . . . father put the treasure,' . . . . . another son. The three . . . . . dug all the fields . . . . . several years, but they . . . . . no treasure. However, many . . . . . grew in the fields . . . . . the sons had dug. . . . . vegetables made them very . . . . . .

Example 2 (Advanced)

It is estimated that in the last two thousand years the world has lost more than a hundred species of animals. A similar number of species of birds has also become extinct. The real significance of . . . . . figures, however, lies in . . . . . fact that almost three-quarters . . . . . all the losses occurred . . . . . the past hundred years . . . . . as a direct result . . . . . man's activities on this . . . . . It is essential for . . . . . whole process of evolution . . . . . the extinction of certain . . . . . should occur over a . . . . . of time. But extinction . . . . . occur by nature's design . . . . . not as a result . . . . . the activities of man . . . . . is by no means . . . . . to the preservation of . . . . . species of animal and . . . . . life.

Conservation means the ..... of a healthy environment ..... a whole. If conservation ..... ignored, then within a ..... short time our water ..... will be found inadequate, ... seas and rivers will ..... fewer fish, our land ..... produce fewer crops, and ..... air we breathe will ..... poisonous. It becomes only ..... matter of time before ..... health deteriorates and before ....., together with every other ..... thing, disappears from the ..... of the earth.

8.10 Open-ended and miscellaneous items

The term 'open-ended' is used to refer to those questions which elicit a completely subjective response on the part of the testees. The response required may range from a one-word answer to one or two sentences:

(One-word answer)
Give the name of the town where the writer had a bad accident.

(Answer in a few words)
You have a friend who is keen on cross-country running. Which event can he enter at the end of the month?

(Sentence answer)
According to the article, why do you think so few foreign cars have been imported into Singapore recently?

When marking open-ended items which require answers in sentences, it is frequently advisable to award at least two or three marks for each correct answer. If the maximum for a correct answer is three marks, for example, the marking guide might be as follows:

Correct answer in a grammatically correct sentence or a sentence containing only a minor error 3
Correct answer in a sentence containing one or two minor errors (but causing no difficulty in understanding) 2
Correct answer but very difficult to understand because of one or more major grammatical errors 1
Incorrect answer in a sentence with or without errors 0

It is always useful to write down precisely how marks should be awarded, even if only one person is marking the items. This marking scheme will then serve as a reminder at all times. It is, of course, essential to write brief guidelines if more than one examiner is marking the items.

Finally, when constructing a reading comprehension test, the test writer should remember to let the text itself determine the types of items set. Indeed, if this principle is observed, it follows that different parts of a particular text will frequently require different item types. It is now becoming common practice for more than one type of item to be used to test comprehension of the same text. Thus, a reading comprehension passage may be followed by one or two multiple-choice items, several true/false items, a few completion items and one or two open-ended items.

8.11 Cursory reading

The title of this section serves as a general term to denote the skills involved in reading quickly, skimming and scanning. The term *skimming* is used to denote the method of glancing through a text in order to become familiar with the gist of the content; *scanning* refers to the skills used when reading in order to locate specific information.

In tests of reading speed the students are generally given a limited time in which to read the text. Care must be taken to avoid constructing questions on the less relevant points in the text, but the students should be expected to be familiar with the successive stages in which the text is developed. The actual reading speed considered necessary will be largely determined by the type of text being read and will vary according to the purpose for which it is being read. It is sufficient to note here that poor readers (native speakers) generally read below 200 words per minute; a speed of between 200 and 300 words per minute is considered to be an average speed; and 300 to 500 words is considered fast. On the whole, it is realistic to expect no more than a reading speed of 300 words from many advanced learners of a second language. Most people tend moreover to read at a slower rate under test conditions or in any situation in which they are required to answer questions on a text.

In tests of skimming, the rubrics generally instruct the students to glance through the text and to note the broad gist of the contents. They are then given a small number of questions concerning only the major points and general outline of the text. Sometimes at the end of the skimming the students are allowed a few minutes to jot down any notes they wish to make, but they are *not* usually allowed to refer back to the text. If the students are allowed to retain the text, the time for answering the various questions on it will be limited in order to discourage them from referring back too often to the text.

In scanning tests, the questions are given to the students before they begin to read the text, thus directing them to read the text for specific information. In such cases, it is helpful to set simple open-ended questions (e.g. 'What is the writer's view of modern transportation?') rather than multiple-choice items. The latter type of item tends only to confuse the students since they then find it necessary to keep in mind four or five options for each item while they are reading.

Tests of speed reading should be administered only when students have been adequately prepared for the tasks involved in such tests. It is grossly unfair to test those reading strategies which have never previously been practised.

Notes and references

1 Based in part on Munby, J 1978 *Communicative Syllabus Design*. Cambridge University Press

2 University of Oxford Delegacy of Local Examinations: *Examination in English Studied as a Foreign Language*, Preliminary Level, 1981

3 Joint Matriculation Board: *Test in English (Overseas)*, March 1982

4 Joint Matriculation Board: *Test in English (Overseas)*, March 1983

5 The text is slightly adapted from an article by Anthony Tucker in *The Guardian*, September 5th 1969.

6 This comprehension test was first constructed by the author for the *Hong Kong English School Certificate Examination* (Education Department, Hong Kong) in 1968, but was later expanded to test awareness of reference devices.

7 Royal Society of Arts: *The Communicative Use of English as a Foreign Language*, Test of Reading, Advanced Level, June 1984

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

9

Testing the writing skills

9.1 The writing skills

The writing skills are complex and sometimes difficult to teach, requiring mastery not only of grammatical and rhetorical devices but also of conceptual and judgemental elements. The following analysis attempts to group the many and varied skills necessary for writing good prose into five general components or main areas.

- language use: the ability to write correct and appropriate sentences;
- mechanical skills: the ability to use correctly those conventions peculiar to the written language – e.g. punctuation, spelling;
- treatment of content: the ability to think creatively and develop thoughts, excluding all irrelevant information;
- stylistic skills: the ability to manipulate sentences and paragraphs, and use language effectively;
- judgement skills: the ability to write in an appropriate manner for a particular purpose with a particular audience in mind, together with an ability to select, organise and order relevant information.

The actual writing conventions which it is necessary for the student to master relate chiefly (at the elementary stages) to punctuation and spelling. However, in punctuation there are many areas in which personal judgements are required, and tests of punctuation must guard against being too rigid by recognising that several answers may be correct. Unfortunately, tests of punctuation and spelling have often tended to inhibit writing and creativity.

Of far greater importance in the teaching and testing of writing are those skills involving the use of judgement. The ability to write for a particular audience using the most appropriate kind of language is essential for both native-speaking and foreign student alike. The use of correct registers becomes an important skill at advanced levels of writing. Failure to use the correct register frequently results in incongruities and embarrassment. Whereas native speakers learn to make distinctions of register intuitively, students of foreign languages usually experience problems in mastering this complex area of language. The various kinds of register include colloquialisms, slang, jargon, archaic words, legal language, standard English, business English, the language used by educated writers of English, etc. The purpose of writing will also help to establish a particular register: for example, is the student writing to entertain, inform, or explain?

A piece of continuous writing may take the form of a narrative, description, survey, record, report, discussion, or argument. In addition to the subject and the format, the actual audience (e.g. an examiner, a teacher, a student, a friend) will also determine which of the various registers is to be used. Consequently, the use of appropriate register in writing implies an awareness not only of a writing goal but also of a particular audience.

After the purpose of writing and the nature of the audience have been established, judgement is again required to determine the selection of the material which is most relevant to the task at hand (bearing in mind the time available). Organisation and ordering skills then follow selection.

9.2 Testing composition writing

An attempt should be made to determine the types of writing tasks with which the students are confronted every day. Such tasks will probably be associated with the writing requirements imposed by the other subjects being studied at school if the medium of instruction is English. Short articles, instructions and accounts of experiments will probably form the main body of writing. If the medium of instruction is not English, the written work will often take the form of consolidation or extension of the work done in the classroom. In both cases, the students may be required to keep a diary, produce a magazine and to write both formal and informal letters. The concern of students following a profession or in business will be chiefly with report-writing and letter-writing, while at college or university level they will usually be required to write (technical) reports and papers.

One large public examining body[1] explicitly states the kinds of writing tasks its examinations test and the standards of writing expected in the performance of those tasks:

> A successful candidate will have passed an examination designed to test ability to produce a selection of the following types of writing:
>
> **Basic Level**: Letter; Postcard; Diary entry; Forms
> **Intermediate Level**: As Basic Level, plus Guide; Set of instructions
> **Advanced Level**: As Intermediate Level, plus Newspaper report; Notes
>
> The candidate's performance will have met the following minimum criteria:
>
> **Basic Level**: No confusing errors of grammar or vocabulary; a piece of writing legible and readily intelligible; able to produce simple unsophisticated sentences.
>
> **Intermediate Level**: Accurate grammar, vocabulary and spelling, though possibly with some mistakes which do not destroy communication; handwriting generally legible; expression clear and appropriate, using a fair range of language; able to link themes and points coherently.
>
> **Advanced Level**: Extremely high standards of grammar, vocabulary and spelling; easily legible handwriting; no obvious limitations on range of language candidate is able to use accurately and appropriately; ability to produce organised, coherent writing, displaying considerable sophistication.

In the construction of class tests, it is important for the test writer to find out how composition is tested in the first language. Although the emphasis in the teaching and testing of the skills in English as a

foreign/second language will of necessity be quite different to the development of the skills in the first language, a comparison of the abilities acquired and methods used is very helpful. It is clearly ludicrous, for instance, to expect in a foreign language those skills which the students do not possess in their own language.

In the past, test writers have been too ambitious and unrealistic in their expectations of testees' performances in composition writing: hence the constant complaint that relatively few foreign learners of English attain a satisfactory level in English composition. Furthermore, the backwash effect of examinations involving composition writing has been unfortunate: teachers have too often anticipated examination requirements by beginning free composition work far too early in the course. They have 'progressed' from controlled composition to free composition too early, before the basic writing skills have been acquired.

However, once the students are ready to write free compositions on carefully chosen realistic topics, then composition writing can be a useful testing tool. It provides the students with an opportunity to demonstrate their ability to organise language material, using their own words and ideas, and to *communicate*. In this way, composition tests provide a degree of motivation which many objective-type tests fail to provide.

In the composition test the students should be presented with a clearly defined problem which motivates them to write. The writing task should be such that it ensures they have something to say and a purpose for saying it. They should also have an audience in mind when they write. How often in real-life situations do people begin to write when they have nothing to write, no purpose in writing and no audience in mind? Thus, whenever possible, meaningful situations should be given in composition tests. For example, a brief description of a real-life situation might be given when requiring the students to write a letter:

> Your pen-friend is going to visit your country for a few weeks with her two brothers. Your house is big enough for her to stay with you but there is not enough room for her brothers. There are two hotels near your house but they are very expensive. The third hotel is cheaper, but it is at least five miles away. Write a letter to your pen-friend, explaining the situation.

Composition titles which give the students no guidance as to what is expected of them should be avoided. Examples of poor titles which fail to direct the students' ideas are *A pleasant evening, My best friend, Look before you leap, A good film which I have recently seen.*

With the emphasis on communicative testing, there is a tendency for test writers to set tasks asking the students to write notes and letters in their own role (i.e. without pretending to be someone else). Tasks requiring the students to act the part of another person are often avoided as it is felt they are less realistic and communicative. However, this is usually far from being the case. It is useful to provide the students not only with details about a specific situation but also with details about the particular person they are supposed to be and the people about (or to) whom they are writing. Compare, for example, the two following tasks:

> (a) Write a letter, telling a friend about any interesting school excursion on which you have been.
>
> (b) You have just been on a school excursion to a nearby seaside town.

However, you were not taken to the beach and you had no free time at all to wander round the town. You are very keen on swimming and you also enjoy going to the cinema. Your teacher often tells you that you should study more and not waste your time. On the excursion you visited the law courts, an art gallery and a big museum. It was all very boring apart from one room in the museum containing old-fashioned armour and scenes of battles. You found this room far more interesting than you thought it would be but you didn't talk to your friends or teacher about it. In fact, you were so interested in it that you left a small camera there. Your teacher told you off because you have a reputation for forgetting things. Only your cousin seems to understand you. Write a letter to him, telling him about the excursion.

Although the former task is one which students may conceivably have to perform in real life, the latter task will result in far more realistic and natural letters from the students simply because the specific details make the task more meaningful and purposeful. The detailed description of both the situation and the person involved helps the students to suspend their disbelief and gives the task an immediacy and realism which are essential for its successful completion.

Two or more short compositions usually provide more reliable guides to writing ability than a single composition, enabling the testing of different registers and varieties of language (e.g. a brief, formal report). If the composition test is intended primarily for assessment purposes, it is advisable not to allow students a choice of composition items to be answered. Examination scripts written on the same topic give the marker a common basis for comparison and evaluation. Furthermore, no time will be wasted by the testees in deciding which composition items to answer. If, on the other hand, the composition test forms part of a class progress test and actual assessment is thus of secondary importance, a choice of topics will cater for the interests of each student.

Finally, the whole question of time should be considered when administering tests of writing. While it may be important to impose strict time limits in tests of reading, such constraints may prove harmful in tests of writing, increasing the sense of artificiality and unreality. Moreover, the fact that candidates are expected to produce a finished piece of writing at their very first attempt adds to this sense of unreality. Students should be encouraged to produce preliminary drafts of whatever they write, and this means giving them sufficient time in an examination to do this. Only in this way can writing become a genuine communicative activity.

9.3 Setting the composition

In addition to providing the necessary stimulus and information required for writing, a good topic for a composition determines the register and style to be used in the writing task by presenting the students with a specific situation and context in which to write. Since it is easier to compare different performances when the writing task is determined more exactly, it is possible to obtain a greater degree of reliability in the scoring of compositions based on specific situations. Furthermore, such composition tests have an excellent backwash effect on the teaching and learning preparatory to the examination.

The difficulty in constructing such compositions arises in the writing of the rubrics. On the one hand, if the description of the situation on which the composition is to be based is too long, then the text becomes more

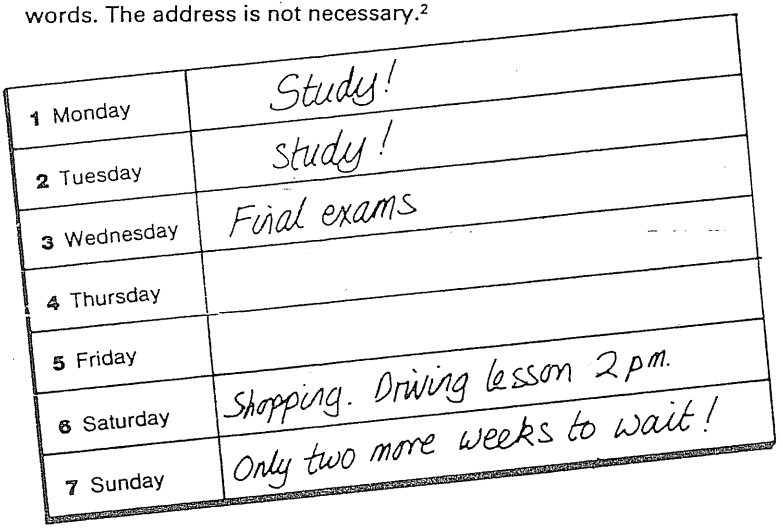
of a reading comprehension test and there will be no common basis for evaluation. On the other hand, however, sufficient information must be conveyed by the rubric in order to provide a realistic, helpful basis for the composition. It is important, therefore, that exactly the right amount of context be provided in simple language written in a concise and lucid manner. The following rubric, for example, can be simplified considerably:

> You have been directed by your superior to compose a letter to a potential client to ascertain whether he might contemplate entering an undertaking that would conceivably be of mutual benefit . . .

The following are provided as examples of situational compositions intended to be used in tests of writing:

Type 1

> Imagine that this is your diary showing some of your activities on certain days. First, fill in your activities for those days which have been left blank. Then, using the information in the diary, write a letter to a friend telling him or her how you are spending your time. Write about 100 words. The address is not necessary.[2]



Type 2

> (Question 1)
> While you are away from home, some American friends are coming to stay in your house. You are leaving before they are due to arrive, so you decide to leave them some notes to help them with all the things they will need to know while staying in the house. Your friends have never been to your country before so there is quite a lot of advice you need to pass on. Write your message on the notelet pad sheet below.
> (*A blank notelet follows.*)
>
> (Question 2)
> While your American friends are staying in your house, they write to say that they are enjoying themselves so much that they would like to spend two weeks visiting some other parts of the country. They would like your advice about what to go and see and where to stay.

Write to your friends giving the best possible advice you can from your own knowledge and experience, with whatever special hints and warnings may be necessary. Make sure your friends know who they can write to for further information of an 'official' kind to help them to plan the best possible holiday.

Write your letter in the space below. It should be between 150 and 200 words in length.[3]
(A blank space follows.)

Type 3

Read the following letter carefully.

176 Wood Lane
London NW2
15th May

Dear Mr Johnson,
I wish to complain about the noise which has come from your home late every night this week. While I realise that you must practise your trumpet some time, I feel you ought to do it at a more suitable time. Ten o'clock in the evening is rather late to start playing. Even if you could play well, the noise would still be unbearable at that time.

I hope that in future you will be a little more considerate of the feelings of others.

Yours sincerely,
W. Robinson

Now write a reply to this letter. You do not play the trumpet but on two or three occasions recently you have played some trumpet music on your record player. You did not play the record very loudly – certainly not as loudly as Mr Robinson's television. You want to tell him this but you do not want to become enemies so you must be reasonably polite in your letter.

Care must be taken in the construction of letter-writing tasks to limit the amount of information to which the student must reply. If this is not done scoring can become extremely difficult.

Type 4 A dialogue can be very useful in providing a basis for composition work. In such a writing task, students must demonstrate their ability to change a text from one register to another, as in the following example:

Read the following conversation carefully.

MR BLACK: What was the weather like while you were camping?
LINDA: Not too bad. It rained the last couple of days, but mostly it was fine. We weren't able to visit the Gorge Waterfalls on the next to the last day, but . . .
MR BLACK: What a pity!
LINDA: Well, apart from that we did everything we wanted to – walking, climbing and just sitting in the sun. We even managed a visit to Hock Cave.
MR BLACK: How on earth did you get that far?
LINDA: We cycled. Oh . . . and we went to the beach quite a few times.
MR BLACK: Did you take your bikes with you?

LINDA: No, we borrowed some from a place in the village.
MR BLACK: Whereabouts were you?
LINDA: Oh, in a lovely valley – lots of woods and about twenty miles from the sea. Just north of Hilson.
MR BLACK: I remember one time when I went camping. We forgot to take a tin-opener!
LINDA: That's nothing. A goat came into our tent in the middle of the night – it ate all the food we had with us!
MR BLACK: Well, you seem to have had a good time.

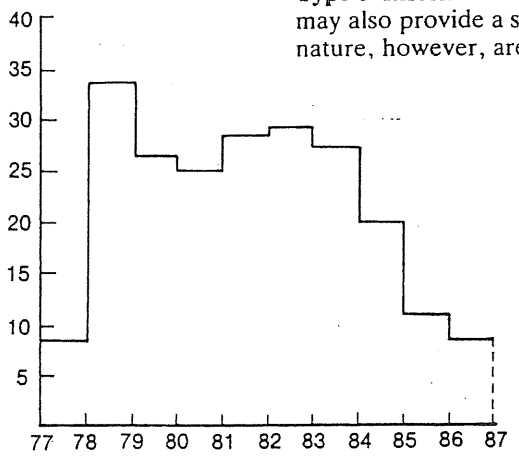
Now write an account of Linda's holiday, using the conversation above as a guide. Imagine other things which happened to her during the camping holiday.

Type 5 Tables containing information are also useful for situational composition since they can generally be read by the students without much difficulty. Moreover, as only a short written text is used, the students are thus not encouraged to reproduce part of the rubric for use in their composition.

Imagine that a local newspaper has asked you to write an article of approximately 250 words about the information in the following table. Write down the conclusions you draw from the figures about the various ways in which people spent their holidays in 1968 as compared with 1988. Attempt to explain the reasons for these differences.

	1968	1988
Travelling abroad	4	17
Going to seaside	38	31
Camping	8	31
Visiting friends/relatives in another town	11	10
Going to another town (but *not* to visit friends/relatives)	16	3
Staying at home	23	8
TOTAL	100	100

Type 6 Information conveyed in the form of a simple graph or histogram may also provide a suitable stimulus for writing. Such writing tasks of this nature, however, are suitable only for more advanced students.



Use the chart together with the information below to give a brief survey of the causes of accidents on Link Road between 1977 and 1987.

1977–78 Road not in great use
1978–79 Nearby road closed: road now in great use
1979–80 Bus stop moved 100 yards
1980–81 No changes
1981–82 Sign: *Beware animals*
1982–83 No parking signs
1983–84 Sign: *No right turn*
1984–85 Double white line: *No overtaking*
(etc.)

Type 7 The stimulus for writing may even take the form of notes.

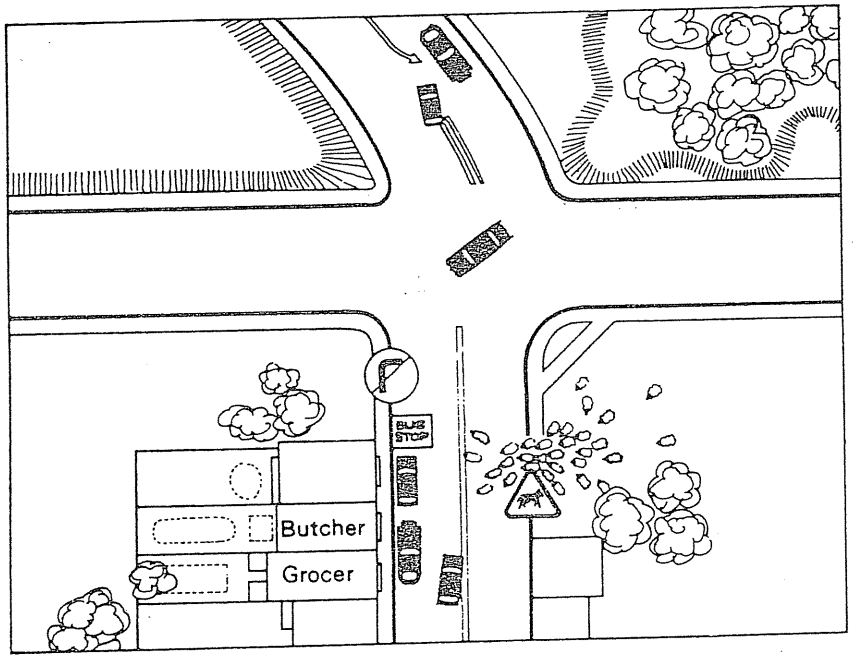
Use the following notes to write an account of an accident for a newspaper. You should write about 250 words.

Cyclist about to turn right.
Not put hand out.
Lorry behind slows down.
Sports car behind lorry overtakes.
Swerves to avoid boy.
Knocks over old man on pavement.

Type 8 An excellent device for providing both a purpose and content for writing is the use of pictures. A picture or series of pictures not only provides the students with the basic material for their composition but stimulates their imaginative powers.

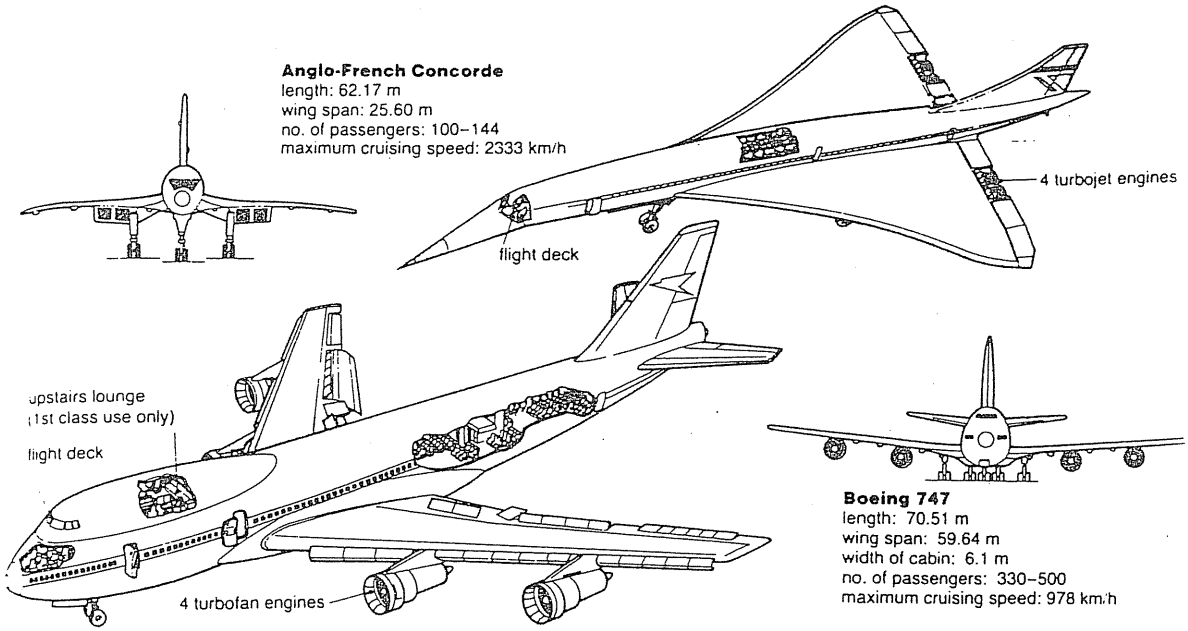
The picture below shows a dangerous junction where accidents often happen. Write a letter to your local newspaper, describing the junction and mentioning some of the dangers and causes of accidents.

Link Road: An Accident 'Black Spot'



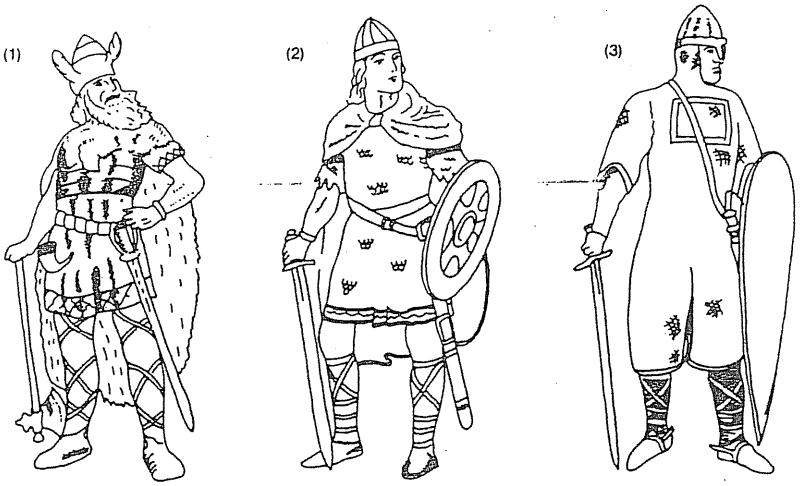
If the stimulus in a situational composition is purely verbal, the testees often tend to reproduce the phrases and sentences contained in it. Pictures and diagrams fortunately avoid this weakness.

Type 9 Pictures and diagrams serving as stimuli for descriptions of objects, people, places and processes may all be used to advantage in a test – as well as those involving the writing of directions, instructions, classifications, comparisons and narratives. The following example[4] shows how the illustration of two aircraft can be used as a basis for a fairly detailed comparison in a test.



Type 10 The following example[5] tests students' ability to describe shapes and linear forms, as it is assumed that none of them will have any knowledge of the technical terms required for reference to each picture. It is a searching test of their descriptive writing ability and is, of course, suitable only for fairly advanced students. The rubric is given for this particular item to help readers to obtain a clearer idea of what is required.

> The pictures below are arranged from the oldest (1) to the most recent (3). Use them to comment on developments in a warrior's clothes and equipment.



9.4 Grading the composition

The chief objection to the inclusion of the composition question as part of any test is generally on grounds of unreliability. Considerable research in the past has shown how extremely unreliable markers are – both in their own inconsistency and in their failure to agree with colleagues on the relative merits of a student's composition.

Markers may award their marks on (i) what a student has written; (ii) what they believe the student meant by what he or she wrote; (iii) handwriting and general appearance of what the student has written; and (iv) previous knowledge of the student. Moreover, two markers may differ enormously in respect of spread of marks, strictness and rank order. For example, marker A may give a wider range of marks than marker B (i.e. ranging from a low mark to a high mark); marker C may have much higher expectations than marker A and thus mark much more strictly, awarding lower marks to all the compositions; and finally marker D may place the compositions in a different order of merit. An example of these differences can be seen in the following table. (The total number of possible marks was 20.)

	Spread		Standard		Order	
	A	B	A	C	A	D
Rick	14	10	14	9	14	9
Amanda	11	9	11	6	11	12
Debbie	10	8	10	5	10	10
Tina	7	7	7	2	7	11
Dave	5	6	5	1	5	6

The whole question of unreliability, however, does not stop here. Even if a student were to take two composition examinations of comparable difficulty, there would be no guarantee whatsoever that he or she would score similar marks on both examinations. This particular type of unreliability is more common to the composition paper than to any other and is termed *test/re-test reliability*. A further complication results from a lack of *mark/re-mark* reliability: in other words, if the same composition is marked by the same examiner on two occasions there may be a difference in the marks awarded.

In spite of all such demonstrations of unreliability, composition is still widely used as a means of measuring the writing skills. The value of practice in continuous or extended writing cannot be stressed too greatly. A student's ability to organise ideas and express them in his or her own words is a skill essential for real-life communication. Thus, composition can be used to provide not only high motivation for writing but also an excellent backwash effect on teaching, provided that the teacher does not anticipate at too early a stage the complex skills required for such a task. Moreover, if a more reliable means of scoring the composition can be used, sampling a student's writing skills in this way will appear a far more valid test than any number of objective tests of grammar.

As is clearly demonstrated at the end of this section, it is impossible to obtain any high degree of reliability by dispensing with the subjective element and attempting to score on an 'objective' basis, according to a carefully constructed system of penalties. However, composition marking can be improved considerably once the subjective element is taken into

account and once methods of reducing the unreliability inherent in the more traditional methods of assessment are employed. To start with, testees should be required to perform the same writing task. Although there may sometimes be a case for a limited choice of composition topics in the classroom, attempts at accurate assessment of writing ability can only be successful if the choice of topic is severely restricted or abolished completely. A well-defined task in terms of one or two situational compositions can help enormously to increase the reliability of the examination.

Because of the inherent unreliability in composition marking, it is essential to compile a banding system – or, at least, a brief description of the various grades of achievement expected to be attained by the class. The following are two examples of descriptions of levels of performance used by a well-known examining body in Britain: table (a) for intermediate-level learners and table (b) for more advanced-level learners.

As with the scoring of oral production, banding systems devised for a particular group of students at a particular level are far preferable to scales drawn up for proficiency tests administered on a national or an international basis.

Table (a)[6]

18–20	Excellent	Natural English, minimal errors, complete realisation of the task set.
16–17	Very good	Good vocabulary and structure, above the simple sentence level. Errors non-basic.
12–15	Good	Simple but accurate realisation of task. Sufficient naturalness, not many errors.
8–11	Pass	Reasonably correct if awkward OR Natural treatment of subject with some serious errors.
5–7	Weak	Vocabulary and grammar inadequate for the task set.
0–4	Very poor	Incoherent. Errors showing lack of basic knowledge of English.

Table (b)[7]

18–20	Excellent	Error-free, substantial and varied material, resourceful and controlled in language and expression.
16–17	Very good	Good realisation of task, ambitious and natural in style.
12–15	Good	Sufficient assurance and freedom from basic error to maintain theme.
8–11	Pass	Clear realisation of task, reasonably correct and natural.
5–7	Weak	Near to pass level in general scope, but with either numerous errors or too elementary or translated in style.
0–4	Very poor	Basic errors, narrowness of vocabulary.

The following rating scale is the result of considerable and careful research conducted in the scoring of compositions in the United States.[7] Only a summary of the scale is shown here and it must be remembered that in its original form slightly fuller notes are given after each item.

Content	
30–27	EXCELLENT TO VERY GOOD: knowledgeable – substantive – etc.
26–22	GOOD TO AVERAGE: some knowledge of subject – adequate range – etc.
21–17	FAIR TO POOR: limited knowledge of subject – little substance – etc.
16–13	VERY POOR: does not show knowledge of subject – non-substantive – etc.
Organization	
20–18	EXCELLENT TO VERY GOOD: fluent expression – ideas clearly stated – etc.
17–14	GOOD TO AVERAGE: somewhat choppy – loosely organized but main ideas stand out – etc.
13–10	FAIR TO POOR: non-fluent – ideas confused or disconnected – etc.
9–7	VERY POOR: does not communicate – no organization – etc.
Vocabulary	
20–18	EXCELLENT TO VERY GOOD: sophisticated range – effective word/idiom choice and usage – etc.
17–14	GOOD TO AVERAGE: adequate range – occasional errors of word/idiom form, choice, usage but meaning not obscured.
13–10	FAIR TO POOR: limited range – frequent errors of word/idiom form, choice, usage – etc.
9–7	VERY POOR: essentially translation – little knowledge of English vocabulary.
Language use	
25–22	EXCELLENT TO VERY GOOD: effective complex constructions – etc.
21–19	GOOD TO AVERAGE: effective but simple constructions – etc.
17–11	FAIR TO POOR: major problems in simple/complex constructions – etc.
10–5	VERY POOR: virtually no mastery of sentence construction rules – etc.
Mechanics	
5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions – etc.
4	GOOD TO AVERAGE: occasional errors of spelling, punctuation – etc.
3	FAIR TO POOR: frequent errors of spelling punctuation, capitalization – etc.
2	VERY POOR: no mastery of conventions – dominated by errors of spelling, punctuation, capitalization, paragraphing – etc.

One point worth noting in the scale is that in each classification (e.g. Content, Organization, Vocabulary, etc.) the lowest grade described as

'Very Poor' ends with the phrase 'OR not enough to evaluate.' (This is not included here.)

Compositions may be scored according to one of two methods: the *impression method* or the *analytic method*. Note, however, that the former method does not involve the use of a rating scale to any large extent.

The **impression method** of marking entails one or more markers awarding a single mark (= multiple marking), based on the total impression of the composition as a whole. As it is possible for a composition to appeal to a certain reader but not to another, it is largely a matter of luck whether or not a single examiner likes a particular script. As has been demonstrated, the examiner's mark is a highly subjective one based on a fallible judgement, affected by fatigue, carelessness, prejudice, etc. However, if assessment is based on several (fallible) judgements, the net result is far more reliable than a mark based on a single judgement.

Generally, three of four markers score each paper, marks being combined or averaged out to reveal the testee's score. The following table shows how four markers can score three compositions using a five-point scale for impression marking.

	Comp. 1	Comp. 2	Comp. 3
Marker A:	3	5	4
Marker B:	2	4	2
Marker C:	2	4	3
Marker D:	3	4	1
Total:	10	17	10(?)
Average:	2.5	4	2.5(?)

In those cases where there is a wide discrepancy in the marks allocated (e.g. Composition 3 in the previous example), the script is examined once again by all four markers and each mark discussed until some agreement is reached. Fortunately, such discrepancies occur only rarely after the initial stages in multiple marking.

All the examiners participating in a multiple-marking procedure are required to glance quickly through their scripts and to award a score for each one. The marking scale adopted may be as little as from 0 to 5 or as large as from 0 to 20 (although it has been the author's experience that most markers prefer to use a 5-point scale or any similar scale with only a few categories in order to obtain a wide range of marks). It is most important that all markers be encouraged to use the whole range of any scale: clearly, marks which bunch around 9 to 12 on a 20-point scale are of little use in discriminating among candidates. It is also important that all the markers read through a certain number of scripts in a given time (usually about 20 per hour) and time themselves. If they find themselves slowing down, marking fewer scripts per hour, they are advised to rest and resume work when they feel able to mark at the required rate. Impression marking is generally found more exhausting than mechanical methods of marking; thus, it is essential that markers stop work when their attention begins to wander or when they find themselves laboriously reading through the content of each composition. Impression marks must be based on impresssion *only*, and the whole object is defeated if examiners start to reconsider marks and analyse compositions. Most examiners find it more

enjoyable than any other method of scoring compositions. Some argue in favour of marking for one or two hours at a stretch in order to maintain consistency, but little conclusive research has been carried out in this area and there appears to be no evidence that marking for a long period produces more consistent and reliable marks than marking for short periods. Impression marking is generally found to be much faster than analytic or mechanical marking. If compositions are scored by three or four impression markers, the total marks have been found to be far more reliable than the marks awarded by one analytic marker. (The comparison is a fair one, since it takes as long for one analytic marker to score a composition as it does four impression markers.) On the other hand, the marks awarded by one impression marker are less reliable than those awarded by one analytic marker.

Since most teachers have little opportunity to enlist the services of two or three colleagues in marking class compositions, the **analytic method** is recommended for such purposes. This method depends on a marking scheme which has been carefully drawn up by the examiner or body of examiners. It consists of an attempt to separate the various features of a composition for scoring purposes. Such a procedure is ideally suited to the classroom situation: because certain features have been graded separately, students are able to see how their particular grade has been obtained. The following is reproduced simply as one example of such an analytic scheme: in this particular case duplicate (blank) copies of this scheme were stencilled by the teacher and attached to the end of each composition.

	5	4	3	2	1
Grammar			X		
Vocabulary				X	
Mechanics		X			
Fluency				X	
Relevance			X		

TOTAL = 14

Note that *Mechanics* refers to punctuation and spelling; *Fluency* to style and ease of communication; and *Relevance* to the content in relation to the task demanded of the student. A 5-point scale has been used.

If the analytic method of scoring is employed, it is essential that flexibility is maintained. At the various levels it may become necessary to change either the divisions themselves or the weighting given to them. At the elementary level, for example, the tester may be far more interested in grammar and vocabulary than in fluency, thus deciding to omit *Fluency*. At the intermediate level, the tester may be particularly interested in relevance and may, therefore, decide to award a maximum of 10 marks for this feature while awarding only 5 marks for each of the others. At the more advanced level, the tester may wish to include separate divisions for organisation and register and to include mechanics and fluency in one division.

A third method of scoring compositions is the **mechanical accuracy** or **error-count method**. Although this is the most objective of all methods of scoring, it is the least valid and is not recommended. The procedure consists of counting the errors made by each testee and deducting the number from a given total: for example, a testee may lose up to 10 marks

for grammatical errors, 5 marks for misuse of words, 5 for misspellings, etc. Since no decision can be reached about the relative importance of most errors, the whole scheme is actually highly subjective. For example, should errors of tense be regarded as more important than certain misspellings or the wrong use of words? Furthermore, as a result of intuition and experience, it is fairly common for an examiner to feel that a composition is worth several marks more or less than the score he or she has awarded and to alter the assessment accordingly. Above all, the mechanical accuracy method unfortunately ignores the real purpose of composition writing – communication; it concentrates only on the negative aspects of the writing task, placing the students in such a position that they cannot write for fear of making mistakes. The consequent effect of such a marking procedure on the learning and teaching of the writing skills can be disastrous.

9.5 Treatment of written errors

Before briefly reviewing some of the attempts to identify error gravity, it is interesting to note a report of an experiment[8] in which native speakers who were not teachers scored written work by its degree of intelligibility rather than by the errors it contained. Native-speaking teachers, on the other hand, evaluated written work by the number and types of errors it contained. Non-native speaking teachers of the language, however, penalised students for what they consider 'basic errors' (e.g. 'He go') and were generally far stricter in their attitude to errors.

Although resulting in a far more positive approach, evaluating written work according to the degree of intelligibility is not always a reliable method of assessment. Frequently, the student's performance – and success in accomplishing the task – may be masked by errors and a tired marker may fail to make the necessary effort to respond to the writing as a means of communication.

An important[9] distinction is now made between *global* and *local* errors. Those errors which cause only minor trouble and confusion in a particular clause or sentence without hindering the reader's comprehension of the sentence are categorised as local errors (e.g. misuse of articles, omission of prepositions, lack of agreement between subject and verb, incorrect position of adverbs, etc.: 'I arrived Leeds.'). Global errors are usually those errors which involve the overall structure of a sentence and result in misunderstanding or even failure to understand the message which is being conveyed (e.g. the misuse of connectives: 'Although the train arrived late, we missed the last bus to the city centre'; the omission, misuse and unnecessary insertion of relative pronouns: 'You should try to be as healthy as the girl arrived on the bicycle a short time ago'; etc.). This useful distinction, which provides criteria for determining the communicative importance of errors, has been further developed recently so that it can be more readily applied to the marking of free writing.[10]

In addition, it is necessary in classroom tests to distinguish between those errors which, though perhaps not resulting in any breakdown in communication, indicate that the student has failed to learn something which has just been taught or which should have been mastered previously.

In most normal writing situations, however, we can only assess what a student writes and *not* what he or she wants to write. For this reason, pictures and diagrams can play a very useful part in testing writing, since they enable the examiner to tell immediately what a student wishes to write. Pictures were recently used by researchers[11] in an experiment to

show how L_2 learners (i.e. less fluent learners) used *avoidance strategies* or *reduction strategies*, avoiding an actual topic. L_1 learners (and possibly more fluent L_2 learners) tended to use *paraphrase strategies* or *achievement strategies*. Fluent performance seems to be characterised by the use of fewer communication strategies of both kinds.

The test writer's attitude to error gravity and approach to treating errors in marking free writing will vary according to the students' level of attainment. At the elementary levels, for example, the test writer will probably be far more tolerant than at the intermediate and advanced levels. At the lower levels he or she will expect more avoidance strategies and more global types of error. What is important at all levels is an awareness of different types of error and of communication strategies, resulting in an increased sensitivity to them.

9.6 Objective tests: mechanics

Punctuation

Type 1 The following type of punctuation item is very popular and is generally used to cover a wide range of punctuation marks. It is not truly objective, and the scoring of such an exercise would take considerable time since punctuation is to a large degree subjective and one particular use of a punctuation mark may well determine the correctness of the punctuation mark following it.

> In the following passage there is no punctuation. Write out the passage, putting in all the punctuation and capital letters.
>
> lend me your pen please peter asked
> i took my pen out of my pocket
> be careful i said
> ill give it back to you in a moment he promised
> dont worry i said you can keep it as long as you want

It is advisable, however, to maintain some degree of control over the task which the testees are expected to perform. One method of doing this is by substituting lines or circles for those punctuation marks which are being tested, thus also facilitating scoring.

Type 2

> Put the correct punctuation mark in each box.
>
> ☐ What do you want, ☐ I asked Henry☐
> ☐ May I use your telephone? ☐ he asked.
> ☐ Certainly ☐☐ I said. ☐ When you ☐ ve finished ☐ please let me know ☐☐
> ☐ I shall only be a moment ☐☐ Henry answered.
> ☐ Has John Lee invited you to his party ☐☐ I asked.
> ☐ No, he hasn ☐ t yet ☐☐ Henry replied.
> ☐ He ☐ s invited Paul ☐ David ☐ Tony and Mary ☐☐ I continued.
> ☐ He ☐ s probably forgotten about me ☐☐ Henry laughed.
> ☐ How strange ☐☐ I answered. ☐ I'm sure he wants you to go to his party. ☐

Type 3 A greater degree of objectivity can be obtained by using the multiple-choice technique, e.g.

Put a circle round the letter (A, B, C, or D) of the correctly punctuated sentence.

A. Tom asked me if I was going to the meeting?
B. Tom asked me, if I was going to the meeting.
C. Tom asked me, 'If I was going to the meeting?'
D. Tom asked me if I was going to the meeting.

Spelling

Type 1: Dictation

As with vocabulary testing, sampling is of primary importance in the construction of spelling tests. Words used in connection with the students' free composition work or everyday writing form the most suitable basis for tests of spelling, although items may also be drawn from the students' reading provided that the tester is aware of the implications of testing the more passive items of the students' vocabulary.

Dictation of long prose passages is still regarded as an essential method of testing spelling. However, dictation measures a complex range of integrated skills and should not be regarded as constituting simply a test of spelling. The dictation of single words, nevertheless, can prove a fairly reliable test of spelling. Several such tests consist of up to fifty words and use similar procedures to the following:

(i) Each word is dictated once by the tester;
(ii) the word is then repeated in a context; and finally,
(iii) the word is repeated on its own.

Type 2: Multiple-choice items

Another fairly widespread method of testing spelling is through the use of multiple-choice items usually containing five options, four of which are spelt correctly. The students are required to select the word which is incorrectly spelt, e.g.

1. A. thief B. belief C. seize D. ceiling E. decieve
2. A. happening B. offering C. occuring D. beginning E. benefiting
3. A. illegal B. generally C. summary D. beggar E. neccessary
4. A. interrupt B. support C. answering D. ocasional E. command

In some tests only four words are given as options, the fifth option being *No mistakes* or *All correct*, e.g.

A. exhibition B. punctually C. pleasure D. obayed E. *All correct*

Type 3: Completion items

Such items as the following differ from similar ones used in tests of vocabulary because sufficient clues are provided both in the blanks and in the definitions to enable the students to know exactly which word is required. The blanks occur only in those parts of the word which give rise to a spelling difficulty for many students. One advantage of such a test is that it does not present the students with incorrect forms. (Many native speakers argue that they frequently fail to recognise correct forms after exposure to misspellings.)

Write one or more letters in each of the spaces in the following words. (The definitions will help you to recognise the correct word.)

1. om s n *something left out*
2. di uade *persuade someone not to do something*
3. o u ing *happening, taking place*
4. rec t *a written statement to show that a bill bas been paid*

It is helpful in tests of this nature to provide a context for the word (in addition to giving a synonym or definition).

1. The horse gal ed (= *ran*) to the front of the race.
2. I doubt if anyone ever prof ed (= *gained*) from that business deal.
3. The school has an enro ment (= *number on its register*) of over 500 students.
4. Don't worry; my dog will go into the water and retr ve (= *bring back*) your ball.

Type 4: Error-recognition items

In these items the students are required to identify (according to its letter) the part of the sentence in which a word has been misspelt.

1. The dissatisfied woman/refused to admit/that there was sufficent/ coffee for everyone.
   A B C D
2. Don't be decieved/by the new procedure:/it hardly differs from/ the old corrupt system.
   A B C D
3. The man was equipped/with a double-barrel shotgun/and his coleague/with an innocent-looking walking-stick.
   A B C D
4. Mr Grant's exaggerated/account of the importance/of his new apointment/ was quite unnecessary.
   A B C D

9.7 Objective tests: style and register

Style

The multiple-choice items below are concerned chiefly with measuring students' sensitivity to style. Some of the distractors in the two examples are incorrect on grammatical grounds while others are grammatically correct but not representative of the kind of English used by an educated native speaker in the particular context in which they appear. Indeed, some test writers distinguish tests of writing from tests of grammar and usage in terms of the performance of native speakers: whereas all native speakers of a language would be expected to score high marks in a test of grammar, only certain educated native speakers possessing the required writing skills would score high marks in an objective test of writing.

Some of the reviews were favourable to the new play

A. and as many were unfavourable.
B. although others of the same amount were unfavourable.
*C. while an equal number were unfavourable.

D. but the same number were unfavourable.
E. in spite of half being unfavourable.

The weather has always been an important factor in people's lives

*A. because of its effects on all aspects of farming.
B. for it has considerable influence over farming.
C. since farmers concern themselves with it.
D. as weather constitutes the dominant worry for farmers.
E. on account of its affecting farming affairs.

(* = *Correct answer*)

Register

The use of the correct register denotes the ability to write for a specific purpose with a specific audience in mind. Confusion and embarrassment result from the use of inappropriate registers. Such tests as the following, however, are not too difficult to construct and present the students with an interesting task, provided that the extract used is written in a fairly distinctive style.

Type 1 The following type of (advanced) register test requires the students to identify those words which are incongruous, replacing each with a much more suitable word. The student is instructed to replace sixteen of the words underlined in the passage.

> It has now been made out beyond any doubt whatsoever that the nicotine contained in tobacco smoke is poisonous. One minute drop of pure nicotine plunged into the bloodstream of a rat is sufficient to kill it. It has also been proved that the nicotine contained in tobacco smoke sends up the pulse rate and the blood pressure. There is also strong evidence that the nicotine content in fags is a primary cause of loss of weight and hungriness. It is also likely that a few heavy smokers will lose control of their finer muscles and be unable to play around with objects with ease and precision. Such a loss of muscle activity may widen the eyes and spoil vision. Moreover, smoking puts back growth in adolescents and lowers athletic ability.
>
> However, the most serious disease connected with smoking is cancer of the lung: the direct connection between smoking and cancer has recently been established so assuredly that cancer research folk and public health authorities throughout the world have begun intensive campaigns against smoking. In certain countries not only are cigarette advertisements banished from cinema and television screens but also makers are forced to print on each packet a warning concerning the dangers of smoking.[12]

Type 2 Matching tests are well-suited to tests of register; such tests can be constructed both at word and sentence level.

(a) Word level: The students are instructed to match each word in List A with a word in List B, depending entirely on how formal or informal a word is – not on its meaning.[13]

List A	*List B*	*Answers*
1. cry	a. boss	(1e)
2. commence	b. gee gee	(2c)
3. kid	c. expire	(3a)
4. pussy	d. hospitalise	(4b)
5. entrain	e. draw	(5d)

(b) Sentence level: The students are instructed to put the letter of the most appropriate sentence in List B with the number of each sentence in List A. The sentences have been taken from instructions, legal documents, scientific English, advertisements, children's comics and newspapers.

List A

1. Build the assembly formers flat on the plan and bend the undercarriage down to the pattern shown.
2. The Tenant shall keep the interior of the premises in good order and condition.
3. A bicycle pump is a device for moving air against a pressure difference.
4. Because the Barcelno has front wheel drive, there's no prop shaft. So you get big car roominess in only thirteen feet.
5. But it's too late! The evil plan, cooked up by the monster Balbo, has led Cato to Madam Zena.
6. Ace driver injured in thrilling race of year.

List B

a. There's a new landmark for lovers and others at Waterloo Station. The Drum Bar and Buffet.
b. An object normally becomes hot when it is placed in the sun.
c. The mixture should be taken three times daily after meals.
d. Gang fight death – youth killed when pushed onto electric line.
e. Any amendment of this certificate or failure to complete any part of it may render it invalid.
f. Give over. I'm not a genius. The radio transmits a kind of buzz. A beam that can be picked up for a couple of miles.

9.8 Controlled writing

There are several ways of controlling students' freedom of expression in their written work and, as a consequence, increasing the reliability of the scoring. However useful such methods are as teaching devices, they will only prove useful for testing purposes if each student is completely familiar with the particular task to be performed: hence the importance of clear instructions followed by at least one example. Sometimes there is even the danger that certain students will feel inhibited rather than helped by such control. Examples of controlled writing exercises are included in this section.

Type 1 The students are given a short reading extract and then required to write a similar paragraph, using the notes they have been given, e.g.

> Although dogs are only animals, they are very useful and help people a lot. For example, certain dogs help farmers to look after their sheep. Some dogs are used for hunting and others help to rescue people. Even now police officers use dogs when they are looking for thieves and criminals. People also teach dogs to race, and dog racing is a sport which many people like. All dogs like eating meat very much and like bones best of all.
>
> Although – horses – animals, – useful – a lot. For example, – horses – people – cattle. Some horses – hunting – pull things. In the past – soldiers – horses – fighting against the enemy. People – horses – horse racing – sport – like. All horses – hay – oats.

Type 2 The following item type[14] is set in a few widely-used examinations and can prove very useful in controlling writing once students are familiar with the conventions observed in the item. Even the following rubric and item may cause difficulty if a student has not previously been given practice in completing such items. Oblique strokes are used in the fragmented sentences chiefly in order to reinforce the impression that the sentences have been given in note form.

> Use the following notes to write complete sentences. Pay careful attention to the verbs underlined and insert all missing words. The oblique lines (/) are used to divide the notes into sections. Words may or may not be missing in each of these sections. Read the example carefully before you start.
>
> Example: Parachute jump from aeroplanes and balloons/be very popular sport/many parts of world.
>
> Parachute jumping from aeroplanes and balloons is a very popular sport in many parts of the world.
>
> Greatest height/from which parachute jump ever make/be over 31,000 metres./1960/doctor in United States Air Force/jump from basket of balloon/and fall nearly 26,000 metres/before open parachute./Fall last 4 minutes and 38 seconds,/and body reach/speed 980 kilometres hour./He land safely in field/13 minutes and 45 seconds/after he jump./On step of basket of balloon/be words/'This be highest step in world.'/When ask/if he like jump again/from such height/doctor shake head.

Type 3 Several types of writing tasks can be based on the following reading extract.[15] Any similar text can also be used for:

- copying with minor alterations: e.g. tense/person changes
- changing the point of view: e.g. *Write this story as seen by . . .*
- changing the style and register: e.g. *Write this story in the form of a newspaper report/a humorous account*, etc.
- adding further information.

> A young man who refused to give his name dived into the river yesterday morning to save a twelve-year-old boy.
>
> The boy, who ran away after being rescued, had been swimming in the river and had caught his foot between two concrete posts under the bridge. He shouted out for help.
>
> At the time, a young man was riding across the bridge on his bicycle. He quickly dismounted and dived fully clothed into the river. He then freed the boy's foot and helped him to the river bank where a small crowd had collected. The boy thanked his rescuer courteously and sincerely, then ran off down the road. He was last seen climbing over a gate before disappearing over the top of the hill.
>
> The young man, who was about twenty years of age, said 'I don't blame the boy for not giving his name. Why should he? If he wants to swim in the river, that's his business. And if I want to help him, that's mine. You can't have my name either!'
>
> He then ran back to the bridge, mounted his bicycle and rode away.
>
> Test (i) Rewrite this story but imagine that you are actually watching everything that is happening. Begin: *There is a small boy swimming . . .*

Test (ii) Rewrite this story as told by (a) the young man who saved the boy and (b) the boy who was saved.

Test (iii) Write this story as if you were giving evidence at a police station.

Test (iv) It was a sunny day but at the time of the rescue it began to rain heavily. Several people were passing nearby on their way to a football match. When the young man went away, everyone thought that he had got wet in the rain. Write out the story, adding these facts.

Type 4 There are also several methods of practising or measuring the ability to link sentences, involving subordination and co-ordination features. Some tasks involve writing up notes in the form of sentences (largely determined by the connectives given). The following is an example of a controlled writing task practising subordination:

Join the short sentences in each of the groups below to form one sentence. Then write each of the finished sentences so as to form a paragraph. Use the joining words given, but note that sometimes no joining word is necessary; also *-ing* denotes the verb ending only.

Each Olympic Games opens. — before
An athlete appears.
He holds a torch. — (-ing)
It has been carried from Mount Olympus in Greece. — which
The ceremony was started in Berlin in 1936. — which
It links the sites of the modern Games with the first Olympic Games.
However, the actual torch ceremony dates back to Ancient Greece. — where
One of the most spectacular events was the torch race. — which
It was always run at night.
The athlete enters the stadium. — When
He is holding the torch. — who
He runs to the huge bowl.
The sacred flame will burn there. — in which

Many such tests do not give the required linkers to the testees but leave them free to join the sentences in whichever way they consider appropriate. Indeed, since such tests are still very subjective and require a lot of time to score, it is often better not to provide the testees with linkers but to leave them free to solve each problem in their own way.

Type 5(a) In some tests of composition, especially at the elementary and intermediate levels, sentences and clauses are provided at first in order to help the students to start writing. They are then required to finish the incomplete sentences in any appropriate way.

Read these sentences. Finish each one and then complete the story in your own words.

One day Hannah and Becky got up early to
go .....
They caught a bus to the large department store
where .....

'Look, that's Pete Shaw over there,' Becky cried.
'Let's .....'
They shouted to Pete but .....
'Why doesn't he look at us?' Hannah asked. He's behaving as
if .....'

(b) This item is similar to the previous one, but here the testees are required to write appropriate sentences rather than clauses. The following example shows how the item type can be used at the upper-intermediate levels.

1. Students who do not know a lot of English can take several steps to prepare for their study in a British university. For example,
   .....
   .....
2. Recent research shows that public opinion is divided on the subject of spending money on defence. About 40 per cent of the country believes we should increase such spending. On the other hand,
   .....
   .....

A well-selected series of such items can include sentences eliciting an ability to use exemplification, contrast, addition, cause, result, purpose, conclusion and summary. Consequently, students can be tested on their ability to use whatever specific functions and notions the test writer wishes.

Type 6 One of the most useful devices for exercising control over the kind of written response required and yet, at the same time, giving the testees considerable freedom of expression is the two-sentence text designed to measure the ability to form a coherent unit of language.[15] For example, testees may be instructed to write a sentence to *precede* the statement:

Moreover, it was impossible to open the windows.

Sample responses could be:

It was very hot in the small room.
There was only one fan in the room, but it was broken.
The door slammed behind John, and he realised he was locked in the room.

In all cases, the students are required to demonstrate an awareness of the communicative nature of language in general and cohesive devices in particular while still retaining a large degree of freedom of response.

Other sample items are:

1. .....
   There was one outside the school entrance, too.
2. .....
   To do this, the water must first be boiled.
3. .....
   These should then be carefully sorted.
4. .....
   For wildlife, however, there are even greater dangers in the pollution of rivers, lakes and seas.
5. .....
   However, there is no reason to be pessimistic.

The communicative nature of this item type would be greatly reduced if the first sentence were given instead of the second sentence. The constraints would then be minimal, the cohesive devices lacking in relevance to a certain degree, and the range of acceptable responses very wide indeed. For example, after the sentence

There was a strange-looking creature outside our door.

Any of the following responses would be acceptable:

I went up to it and stroked it tenderly.
Do you like it?
Mrs Lee screamed.
The next thing I knew I was lying on my back.
There were also several cats and dogs.
~~The telephone suddenly rang!~~
It was a hot day.

Notes and references

1 Royal Society of Arts: *The Communicative Use of English as a Foreign Language*

2 University of Cambridge Local Examinations Syndicate: *Preliminary English Test* (revised version)

3 Royal Society of Arts: *The Communicative Use of English as a Second Language*, Test of Writing: Advanced Level, summer 1984

4 Heaton J B 1986 *Writing through pictures*. Longman

5 Joint Matriculation Board: *Test in English (Overseas)*, March 1983

6 University of Cambridge Local Examinations Syndicate *Cambridge Examinations in English* (Table (a) = *First Certificate in English*; Table (b) = *Certificate of Proficiency in English*)

7 Hartfiel, Faye *et. al.* 1985 ESL Composition Profile. *Learning ESL Composition*. Newbury House

8 Hughes, A and Lascaratou, C 1982 Competing criteria for error gravity. *ELT Journal* **36**(3)

9 Burt, M K and Kiparsky C 1972 *The Gooficon: a repair manual for English*. Newbury House
Hendrickson, J (ed.) 1979 Error Analysis and Error Correction in Language Teaching. *RELC Occasional Papers No.* **10**

10 Tomiyana, M 1980 Grammatical errors and communication breakdown. *TESOL Quarterly* **14**(1)

11 Ellis, R 1984 Communication strategies and the evaluation of communicative performance. *ELT Journal* **38**(1).

12 made out (proved), plunged (injected), sends up (increases), fags (cigarettes), hungriness (appetite), play around with (manipulate), widen (extend), spoil (impair), puts back (retards), lowers (reduces), assuredly (conclusively), folk (organisations), begun (launched), banished (banned), forced (required), makers (manufacturers)

13 In an excellent article 'Style and Register Tests', in *Objektive Tests im Englischunterricht der Schule and Universität*, Athenäum Verlag, Robert Pynsent draws attention to the use of semantic and structural distractors in matching tests of register – i.e. including words with the same meaning (or phrases with identical structures) but in a different register.

14 This type of item has been used in the University of Cambridge Local Examinations Syndicate: *First Certificate in English*.

15 Samonte, A L 1976 Techniques in Teaching Writing. *RELC Journal* **1**(1)
Sharwood-Smith, 1976 New Directions in Teaching Written English. *Forum* **XIV**(2)

10

Criteria and types of tests

10.1 Validity

This section attempts to summarise much of what was contained in Chapter 1. Briefly, the validity of a test is the extent to which it measures what it is supposed to measure *and nothing else*. Every test, whether it be a short, informal classroom test or a public examination, should be as valid as the constructor can make it. The test must aim to provide a true measure of the particular skill which it is intended to measure: to the extent that it measures external knowledge and other skills at the same time, it will not be a valid test. For example, the following test item is invalid if we wish solely to measure writing ability: 'Is photography an art or a science? Discuss.' It is likely to be invalid simply because it demands some knowledge of photography and will consequently favour certain students.

Similarly, many oral interview tests run the risk of assessing personality as well as oral proficiency. The latter case is an interesting one, however, as it can be strongly argued that success should be measured by concentrating on whether the task set has been achieved using whatever strategies the students have at their disposal rather than simply on pure linguistic ability. In other words, why should personality and a variety of useful non-verbal strategies be ignored in assessing oral performance in a language? The answer to this question, however, must always depend on the purpose of the test: an achievement test or a classroom progress test might well exclude such factors as personality while a proficiency test or public examination might consider such factors as important in evaluating oral ability in the target language. Nevertheless, the question serves to illustrate a significant difference between the rather narrow psychometric–structural approach and the broader communicative approach to testing.

Face validity

Even a superficial inspection of the essay item referred to in the first paragraph in this section would be sufficient to reveal that it was not valid. This type of validity, in fact, is often referred to as *face validity*: if a test item looks right to other testers, teachers, moderators, and testees, it can be described as having at least face validity. It is, therefore, often useful to show a test to colleagues and friends. As constructors of the test, we can become so involved in the test that we sometimes fail to stand back and look at the individual test items objectively. Only if the test is examined by other people can some of the absurdities and ambiguities then be discovered.

Language tests which have been designed primarily for one country and are adopted by another country may lack face validity. A vocabulary or reading comprehension test containing such words as 'typhoon', 'sampan', 'abacus', and 'chopsticks' will obviously not be valid in East Africa no matter how valid and useful a test it has proved in Hong Kong. The same argument applies to many standardised tests designed for immigrants in the United States and used later in many other parts of the world.

Although no substitute for empirical validity in public examinations and standardised tests, face validity can provide not only a quick and reasonable guide but also a balance to too great a concern with statistical analysis. Moreover, the students' motivation is maintained if a test has good face validity, for most students will try harder if the test looks sound. If, on the other hand, the test appears to have little of relevance in the eyes of the student, it will clearly lack face validity. Possibly as a direct result, the student will not put maximum effort into performing the tasks set in the test: hence the reliability of the test will be affected.

It is possible for a test to include all the components of a particular teaching programme being followed and yet at the same time lack face validity. For example, a reading test for engineers could have most of the grammatical features of the language of engineering (e.g. frequent use of the passive voice, nominal strings, etc.) as well as most of the language functions and notions associated with reading and writing engineering texts (e.g. defining, classifying, hypothesising, drawing conclusions, describing processes, expressing notions of quantity and amount, etc.). Even so, however, the test will lack face validity if the subject of the reading text concerns, say, public institutions in Britain.

The concept of face validity is far from new in language testing but the emphasis now placed on it is relatively new. In the past, face validity was regarded by many test writers simply as a public relations exercise. Today, however, most designers of communicative tests regard face validity as the most important of all types of test validity. Indeed, many argue that a test must look valid even as far as the reproduction of the material itself is concerned: thus, a test of reading comprehension using such authentic tasks as reading and skimming newspapers must contain actual newspapers – or, at least, articles printed in exactly the same way as they appeared in the newspaper from which they were taken.

Content validity

This kind of validity depends on a careful analysis of the language being tested and of the particular course objectives. The test should be so constructed as to contain a representative sample of the course, the relationship between the test items and the course objectives always being apparent. There is a strong tendency, especially in multiple-choice testing, to test only those areas of the language which lend themselves readily to testing. Many tests of phonology, for instance, concentrated in the past on testing phoneme discrimination rather than the more important features of stress and intonation: one cannot help but suspect the reason for this was simply that phoneme tests were easier to construct than items testing stress and intonation.

When embarking on the construction of a test, the test writer should first draw up a table of test specifications, describing in very clear and precise terms the particular language skills and areas to be included in the test. If the test or sub-test being constructed is a test of grammar, each of the grammatical areas should then be given a percentage weighting (e.g.

the future simple tense 10 per cent, uncountable nouns 15 per cent, relative pronouns 10 per cent, etc.) as touched upon previously. If the test or sub-test concerns reading, then each of the reading sub-skills should be given a weighting in a similar way (e.g. deducing word meanings from contextual clues 20 per cent, search-reading for specific information 30 per cent, reading between the lines and inferring 12 per cent, intensive reading comprehension 40 per cent, etc.). It scarcely matters what the total percentage is: the important point is that the test writer has attempted to quantify and balance the test components, assigning a certain value to indicate the importance of each component in relation to the other components in the test. In this way, the test should achieve content validity and reflect the component skills and areas which the test writer wishes to include in the assessment.

Construct validity

If a test has *construct validity*, it is capable of measuring certain specific characteristics in accordance with a theory of language behaviour and learning. This type of validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills. For example, it can be argued that a speed reading test based on a short comprehension passage is an inadequate measure of reading ability (and thus has low construct validity) unless it is believed that the speed reading of short passages relates closely to the ability to read a book quickly and efficiently and is a proven factor in reading ability. If the assumption is held that systematic language habits are best acquired at the elementary level by means of the structural approach, then a test which emphasises the communicative aspects of the language will have low construct validity. Conversely, if a communicative approach to language teaching and learning has been adopted throughout a course, a test comprising chiefly multiple-choice items will lack construct validity.

Empirical validity

A fourth type of validity is usually referred to as *statistical* or *empirical* validity. This validity is obtained as a result of comparing the results of the test with the results of some criterion measure such as:

- an existing test, known or believed to be valid and *given at the same time*; or
- the teacher's ratings or any other such form of independent assessment *given at the same time*; or
- the *subsequent* performance of the testees on a certain task measured by some valid test; or
- the teacher's ratings or any other such form of independent assessment *given later*.

Results obtained by either of the first two methods above are measures of the test's *concurrent validity* in respect of the particular criterion used. The third and fourth methods estimate the *predictive validity* of a test which is used to predict future success. We could estimate the predictive validity of a language test administered to civil engineers embarking on a civil engineering course given in the medium of English, for instance, if we measured their subsequent performances on their academic courses and compared these results with the results of the language test given at the beginning of their course.

The test situation or the technique used is always an important factor in determining the overall validity of any test. Although an ideal test

situation will by no means guarantee validity, a poor test situation will certainly detract from it. Is a listening comprehension test valid if the students hear only a disembodied voice on, say, a poor quality tape recorder?

Moreover, the fact that a new test compares favourably with another *supposedly valid* test will not necessarily ensure that the new test is valid, particularly if the other test is not really a valid measure, itself. In short, how far can we trust the criteria we use for establishing validity? This is one of the major criticisms of the whole concept of empirical validity made by the communicative school of test writers. The argument is simply that the established criteria for measuring validity are themselves very suspect: two invalid tests do not make a valid test.

10.2 Reliability

Reliability is a necessary characteristic of any good test: for it to be valid at all, a test must first be reliable as a measuring instrument. If the test is administered to the same candidates on different occasions (with no language practice work taking place between these occasions), then, to the extent that it produces differing results, it is not reliable. Reliability measured in this way is commonly referred to as *test/re-test reliability* to distinguish it from *mark/re-mark reliability* and the other kinds of reliability described later in this section. This latter kind of reliability denotes the extent to which the same marks or grades are awarded if the same test papers are marked by (i) two or more different examiners or (ii) the same examiner on different occasions. In short, in order to be reliable, a test must be consistent in its measurements.

Reliability is of primary importance in the use of both public achievement and proficiency tests *and* classroom tests. Methods of estimating the reliability of individual items in a test will be indicated in the next chapter. However, an appreciation of the various factors affecting reliability is important for the teacher at the very outset, since many teachers tend to regard tests as infallible measuring instruments and fail to realise that even the best test is indeed a somewhat imprecise instrument with which to measure language skills.

Factors affecting the reliability of a test are:

- the extent of the sample of material selected for testing: whereas validity is concerned chiefly with the content of the sample, reliability is concerned with the size. The larger the sample (i.e. the more tasks the testees have to perform), the greater the probability that the test as a whole is reliable – hence the favouring of objective tests, which allow for a wide field to be covered.
- the administration of the test: is the same test administered to different groups under different conditions or at different times? Clearly, this is an important factor in deciding reliability, especially in tests of oral production and listening comprehension.

The way in which this factor differs from test situation validity can be seen from the following example: if a recording for a listening comprehension test is initially poor in quality, then it is poor in quality for all testees. This will consequently make for invalidity (unless speech has been *deliberately* masked with noise, as a testing device). But if the quality of the recording is good and if certain groups hear it played under good acoustic conditions while other groups hear it under poor acoustic conditions, this will make for unreliability and therefore invalidity.

– test instructions: are the various tasks expected from the testees made clear to *all* candidates in the rubrics?
– personal factors such as motivation and illness.
– scoring the test: one of the most important factors affecting reliability. Objective tests overcome this problem of marker reliability, but subjective tests are sometimes faced with it: hence the importance of the work carried out in the fields of the multiple-marking of compositions and in the use of rating scales.

Communicative language testing has recently introduced another dimension to the whole concept of reliability: namely, profile reporting. In order to obtain a full profile of a student's ability in the target language, it is necessary to assess his or her performance separately for each of the different areas of communication: e.g. listening comprehension, speaking and listening, reading, reading and writing (summarising, etc.), and writing. Furthermore, performance is assessed according to the purpose for which the language is to be used: e.g. academic, occupational, social survival. The object of the sub-tests through which performance is assessed is to indicate the extent of the learner's mastery of the various language skills which he or she will require for a particular purpose. A score or grade is given for each of the skills or areas selected for testing, and an average mark is eventually obtained. This latter mark, however, is only given alongside the various scores which have contributed to it. Profile reporting is thus very valuable for placement purposes, and indeed it is an essential feature of one of the most widely used proficiency tests set in Britain and administered in many countries throughout the world.[1] A student's performance on the various parts of the test can be shown in the form of a simple table or chart, in which the target score appears beside the student's score. It is thus very easy to compare a student's performance levels in each area with the required levels.

One method of measuring the reliability of a test is to re-administer the same test after a lapse of time. It is assumed that all candidates have been treated in the same way in the interval – that they have either all been taught or that none of them have. Provided that such assumptions (which are frequently hard to justify) can be made, comparison of the two results would then show how reliable the test has proved. Clearly, this method is often impracticable and, in any case, a frequent use of it is not to be recommended, since certain students will benefit more than others by a familiarity with the type and format of the test. Moreover, in addition to changes in performance resulting from the memory factor, personal factors such as motivation and differential maturation will also account for differences in the performances of certain students.

Another means of estimating the reliability of a test is by administering parallel forms of the test to the same group. This assumes that two similar versions of a particular test can be constructed: such tests must be identical in the nature of their sampling, difficulty, length, rubrics, etc. Only after a full statistical analysis of the tests and all the items contained in them can the tests safely be regarded as parallel. If the correlation between the two tests is high (i.e. if the results derived from the two tests correspond closely to each other), then the tests can be termed reliable.

The split-half method is yet another means of measuring test reliability. This method estimates a different kind of reliability from that

estimated by test/re-test procedures. The split-half method is based on the principle that, if an accurate measuring instrument were broken into two equal parts, the measurements obtained with one part would correspond exactly to those obtained with the other. The test is divided into two and the corresponding scores obtained, the extent to which they correlate with each other governing the reliability of the test as a whole. One procedure widely used is to ascertain the correlation between the scores on the odd numbered items and those on the even numbered items. However, if the items are graded according to increasing difficulty, division according to odd and even numbers would not be an accurate means of assessing reliability, since item 2 would be slightly more difficult than item 1, item 4 again more difficult than item 3, and so on. A more accurate procedure is to balance the items as follows:

item	1	4	5	8	9	12
against item	2	3	6	7	10	11

However, it would be better, though less convenient, to allow chance to decide which items go into one half and which into the other.

The reliability of the whole test can be estimated by using the formula:

$$r_{11} = \frac{N}{N-1}\left(1 - \frac{m(N-m)}{Nx^2}\right)$$

where N = the number of items in the test;
m = the mean score on the test for all the testees (see page 175);
x = the standard deviation of all the testees' scores (see page 176), and
r_{11} = reliability.

(Note that in this formula, x has to be squared.)

In Sections 11.1 and 11.2 the calculation of the mean and standard deviation of scores on a language test containing 40 items is illustrated. The mean is found to be 27 and the standard deviation 4.077. Using these figures with the above formula, we obtain:

$$r_{11} = \frac{40}{39}\left(1 - \frac{27 \times 13}{40 \times 16.662}\right) = 0.484$$

This formula is simple to use since (i) it avoids troublesome correlations and (ii), in addition to the number of items in the test, it involves only the test mean and standard deviation, both of which are normally calculated anyhow as a matter of routine.

Finally, it should be noted that a test can be reliable without necessarily possessing validity. However, reliability is clearly inadequate by itself if a test does not succeed in measuring what it is supposed to measure.

10.3 Reliability versus validity

As we have seen, test validity and reliability constitute the two chief criteria for evaluating any test, whatever the theoretical assumptions underlying the test. The fundamental problem, however, lies in the conflict between reliability and validity. The ideal test should, of course, be both reliable and valid. However, the greater the reliability of a test, the less validity it usually has. Thus, the real-life tasks contained in such productive skills tests as the oral interview, role-play, letter writing, etc. may have been given high construct and face validity at the expense of reliability.

Objective tests are clearly not subject to such a degree of unreliability. But does this mean that all forms of subjective testing should be abandoned in favour of objective testing? For most purposes objective tests are vastly inferior to such more meaningful and communicative tasks as free-writing, role playing and problem-solving. Language-learning behaviour cannot be demonstrated solely by means of an ability to select correct options from given alternatives. Language use simply does not function in this way.

The choice facing the test writer is, therefore, whether to attempt to increase the validity of a test known to be reliable or else to increase the reliability of a test known to be valid. If the test writer tries to do the former, he or she will be faced with an impossible task because the very features which make the test reliable (usually multiple-choice and completion items tested out of context) are those very features which render the test invalid. Consequently, it is essential to devise a valid test first of all and then to establish ways of increasing its reliability.

One effective way of increasing test reliability in such cases is by means of a carefully drawn up banding system or rating scale. Such a scale (with a clear and concise description of the various characteristics of performance at each level) enables the marker to identify precisely what he or she expects for each band and then assign the most appropriate grade to the task being assessed. Furthermore, markers are encouraged in this way not only to be consistent in their marking but also to formulate judgements in qualitative terms before later converting such judgements into quantitative assessments. An example of such a rating scale is given in the previous chapter: it is sufficient to emphasise here that reliability can be increased by means of profile reporting and qualitative judgements (rather than quantitative ones).

10.4 Discrimination

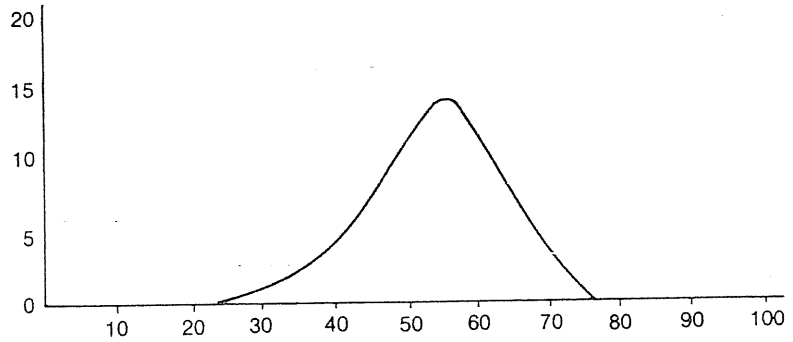
Sometimes an important feature of a test is its capacity to discriminate among the different candidates and to reflect the differences in the performances of the individuals in the group. For example, 70 per cent means nothing at all unless all the other scores obtained in the test are known. Furthermore, tests on which almost all the candidates score 70 per cent clearly fail to discriminate between the various students. Tests which are designed for a large test population (and which are to be standardised) are first tried out on a representative sample of students. This small sample mirrors the much larger group for whom the test is intended. The results of the test are then examined to determine the extent to which it discriminates between individuals who are different. When the final version of the test is eventually used, therefore, its discriminatory powers have already been established. Consequently, there will then be little need for concern if, for example, it is found that the scores of individuals in a group cluster around a central point. The test has been proved capable of discriminating; it does not do so in this case because there is nothing to discriminate.

The extent of the need to discriminate will vary depending on the purpose of the test: in many classroom tests, for example, the teacher will be much more concerned with finding out how well the students have mastered the syllabus and will hope for a cluster of marks around the 80 per cent and 90 per cent brackets. Nevertheless, there may be occurrences in which the teacher may require a test to discriminate to some degree in order to assess relative abilities and locate areas of difficulty.

Even the best test can never be so precise as to determine the true score of a testee to within 2 or 3 per cent of a certain mark. This lack of precision is often referred to as the *margin of error* in a test. It is important to be aware of this grey, borderline area in test scores, especially when cut-off points (i.e. pass/fail levels) are being fixed. Ironically, such cut-off points are usually fixed around a middle range of scores (i.e. 40 per cent or 50 per cent), affecting most of the students. It is precisely in this range of scores where the important decisions about pass or fail are usually made.

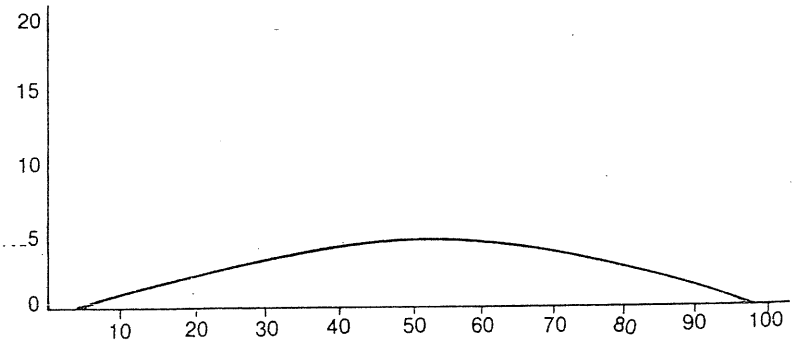
There is a limit to the extent to which such scores can be refined in order to enable these pass/fail decisions to be made with complete certainty. In other words, the grey borderline area referred to in the previous paragraph will always remain. The situation can be improved considerably, however, if the scores themselves can be spread out over the whole range of the scale. In this way, fewer students will be affected when discriminations are made between scores on the critical part of the scale.

In the following graph showing the usual pattern of scores, the examiner will be uncertain about the large number of students whose scores fall within 2 per cent on either side of the pass/fail score of 50 per cent.



If the test can be constructed so as to discriminate as much as possible, the scores will then be spread out over a far wider range as shown on the following graph. It will still be impossible to be certain when distinguishing between a student who has scored 49 per cent (and thus failed) and a student who has scored 51 per cent (and passed). However, the number of students in this borderline area will now be much smaller.

The spread of scores will now appear in the graph like this:



How can test scores be spread in this way and the test consequently make finer discriminations among the testees? Briefly, the items in the test should be spread over a wide difficulty level as follows:

- extremely easy items
- very easy items
- easy items
- fairly easy items
- items below average difficulty level
- items of average difficulty level
- items above average difficulty level
- fairly difficult items
- difficult items
- very difficult items
- extremely difficult items.

10.5 Administration

A test must be practicable: in other words, it must be fairly straightforward to administer. It is only too easy to become so absorbed in the actual construction of test items that the most obvious practical considerations concerning the test are overlooked. The length of time available for the administration of the test is frequently misjudged even by experienced test writers, especially if the complete test consists of a number of sub-tests. In such cases sufficient time may not be allowed for the administration of the test, the collection of the answer sheets, the reading of the test instructions, etc. In the case of all large-scale tests, the time to be allowed should be decided on as a result of a pilot administration of the test (i.e. a tryout of the test to a small but representative group of testees).

Another practical consideration concerns the answer sheets and the stationery used. Many tests require the testees to enter their answers on the actual question paper (e.g. circling the letter of the correct option), thereby unfortunately reducing the speed of the scoring and preventing the question paper from being used a second time. In some tests, the candidates are presented with a separate answer sheet, but too often insufficient thought has been given to possible errors arising from the (mental) transfer of the answer from the context of the item on the question paper to the answer sheet itself. Confusion may result, for example, if the items are numbered vertically on the question papers and horizontal numbering is adopted for the corresponding answer sheet:

1. You'd already left by seven o'clock, ... you?
   A. didn't
   B. weren't
   C. hadn't
   D. haven't
2. If you take swimming lessons, you . . . . . soon.
   A. will be able to swim
   B. swim
   C. can swim
   D. shall have swum

3. Did anyone tell Tim ..... the careless error he made?
   A. off
   B. over
   C. on
   D. about

Put a cross (X) in the box containing the letter of the correct answer.

1. [A|B|C|D] 2. [A|B|C|D] 3. [A|B|C|D]

The use of separate answer sheets, however, greatly facilitates marking (through the use of a mask or key) and is strongly recommended when large numbers of students are being tested.

It is of paramount importance that examiners are fully conversant with the test situation. If the test is to be administered by several examiners working in different test centres, clear directions specifying exactly what each examiner should say and do should be issued in order to ensure that exactly the same procedure is followed in each centre. Although this principle seems obvious enough, it is extremely difficult in many cases for test writers to see their own test from the point of view of the people conducting the test, simply because they are so closely involved in their test that they are inclined to take too many things for granted. Consequently, wherever possible, all test arrangements should be discussed in detail by the test writer and by those conducting the test, the various steps in the administering of each sub-test being stated in simple language and clearly numbered. This is particularly essential in tests of listening comprehension and oral production, where the administrator's role in the test is so important. Accompanying these instructions should be a clear statement of aims together with a comprehensive (but simple) marking scheme.

Before beginning to construct a test, the test writer must make certain that the necessary equipment will be available in each centre and that there will be a high degree of standardisation in the test administration. Clearly, it is useless to record talks or dialogues on tape if certain test centres do not have a tape recorder. What is not so obvious, however, is the potential unreliability of a listening test resulting from the different sizes of the rooms where the test is administered and the different degrees of interference caused by extraneous noise. The question of practicability, however, is not confined solely to aural/oral tests: such written tests as situational composition and controlled writing tests depend not only on the availability of qualified markers who can make valid judgements concerning the use of language, etc., but also on the length of time available for the scoring of the test.

A final point concerns the presentation of the test paper itself. Where possible, it should be printed or typewritten and appear neat, tidy and aesthetically pleasing. Nothing is worse and more disconcerting to the testee than an untidy test paper, full of misspellings, omissions and corrections.

10.6 Test instructions to the candidate

Since most students taking any test are working under certain mental pressures, it is essential that all instructions are clearly written and that examples are given. Unless all students are able to follow the instructions, the test will be neither reliable nor valid. Grammatical terminology should be avoided and such rubrics as the following rewritten:

> Put the correct pronoun in the blanks.
> Choose one of the following verbs to go in each blank space and put it in the correct tense.

Students may be able to perform all the required tasks without having any knowledge of formal grammar. Indeed, since their knowledge of formal grammar is not being tested, all reference to grammatical terms should be avoided. Thus, the first of the rubrics above should be rewritten and the phrase 'words like the following' (followed by examples) be used to replace 'pronouns'; the second rubric should refer to 'words' instead of verbs, and examples should be given so that students are shown the tense changes they are required to make. This principle does not apply only to grammatical terms: if students are instructed to put a tick opposite the correct answer, an example of what is meant by the word 'tick' should be given – e.g. (✓). The same applies to crosses, circles, underlining, etc.

Sometimes it is difficult to avoid writing clumsy rubrics or rubrics consisting of complex sentences above the difficulty level being tested, e.g.

> Answer each of the following questions by selecting the word or group of words which best completes each sentence from the words or groups of words which are lettered A, B, C, and D.
>
> For each of the blanks in the following sentences choose one of the words or groups of words which best completes the sentence. Write in the space shown by the dotted line the letter corresponding to the word, or group of words, which best completes the sentence.

One possible solution is to write the rubric in short sentences – clearly and concisely:

> Choose the word or group of words which best completes each sentence. Each is lettered A, B, C, or D. Put a circle round the letter of the correct answer.

Another solution is to use the students' first language when the test group is monolingual. However, this procedure is recommended only at the elementary level where absolutely necessary.

The rubrics in too many existing tests assume that the testees already know what to do. While this can be excused to a certain extent in class tests where the teacher is present to explain, it is very disturbing when it occurs in more widely used achievement and proficiency tests where no such help is available. However, it is often difficult to strike the right balance between short, clear instructions and long, involved rubrics. A rubric should never, in itself, become a test of reading comprehension. The following are further examples of clear instructions:

> Complete each of the following sentences. Write in the blank space the letter of the correct word or words. The first one is an example.
>
> There are 40 questions in this section. Each question consists of an incomplete sentence. This incomplete sentence is followed by five possible ways of completing it. Each way is labelled A, B, C, D, or E. Choose the answer which you think best completes the sentence. The first one is done for you.

Each word should be carefully considered in the writing of a rubric. For example, the word 'best' is used in certain instances instead of 'correct' because many test items (especially those in tests of vocabulary) contain

several 'correct' answers, although only one is really acceptable and clearly the required answer. Finally, all rubrics should be tried out on pilot groups in the same way in which items are tried out.

Because the writing of clear and concise rubrics is so difficult, it is essential that simple examples are provided for the testees. One or two examples of the type of task set are recommended, unless the particular testing technique adopted is one with which testees are extremely familiar. Indeed, if the testees are all unfamiliar with the type of test being given, it is advisable for the test supervisor to work through a few examples with them. In one test of proficiency, the testees are given five 'practice' items to work through, their answers being subsequently checked before the test is commenced. This small precaution ensures that all testees are conversant with the type of test in which they are about to participate. In certain other tests, a 'practice' test is administered beforehand to the testees. Such a test is specially constructed to include examples of all the types of items in the test paper.

If new testing techniques are being used on a large scale for the first time, it is essential for the test writers and test administrators concerned to construct sample items and rubrics to send to schools well in advance, together with sufficient detailed information about the new procedures. No test can possibly be valid if the techniques adopted are so new and unfamiliar as to bewilder the testees.

10.7 Backwash effects

In Chapter 1 and throughout this book the importance of the backwash effects of testing on teaching has been emphasised. In Chapter 1 reference was made to oral examining, where it was pointed out that, in spite of possible unreliability, oral tests should be continued as far as possible in certain language learning situations if for no other reason than the backwash effects they have on the teaching that takes place before the test. The possible consequences of many reading comprehension tests on the development of the reading skills were cited as another example of the backwash effects of testing. Each element and skill has been treated in the book in relation to its potential influence on teaching. Nevertheless, the importance of the influence of testing on teaching is worth emphasising again, since the test constructor can so easily become too deeply involved in test statistics and in other test criteria.

A larger issue at stake is the effect of objective tests on language learning in general. Important questions are raised; many as yet remain unanswered. For example, do objective tests frequently lead to a greater emphasis on accuracy than on fluency? It is highly possible that, in spite of all efforts, many testing techniques still emphasise the negative aspects of language learning, encouraging teachers and students to place more emphasis on correctness than on the communicative aspects of language learning. A number of objective tests also encourage the teaching of language in artificially constructed situations, thereby reducing motivation in language learning.

Other issues are equally important in their implications. How much influence do certain tests exert on the compilation of syllabuses and language teaching programmes? How far is such an influence harmful or actually desirable in certain situations? Again, what part does coaching play in the test situation? Is it possible to teach effectively by relying solely on some of the techniques used for testing? Clearly, the answers to these

and other questions remain in some doubt. All that can be done at present is to discourage as actively as possible the use of testing techniques as the chief means of practising certain skills. While coaching undoubtedly plays a part in increasing test scores, good teaching can do far more. Moreover, provided that the students have at least one opportunity to participate in a practice test paper before embarking on a particular test, coaching by itself will produce little improvement in test results.

Consequently, while we may deplore, and must guard against, certain backwash effects of testing on the one hand, it is fair to point out on the other hand that testing has been one of the greatest single beneficial forces in changing the direction of language teaching in many areas and in encouraging the more responsive teachers to examine not only their own teaching methods but also the language they are teaching.

10.8 Types of tests

There is some confusion regarding the terminology used to denote the different types of language tests in use. Most test specialists, however, agree on the following broad divisions: achievement/attainment tests, proficiency tests, aptitude tests and diagnostic tests.

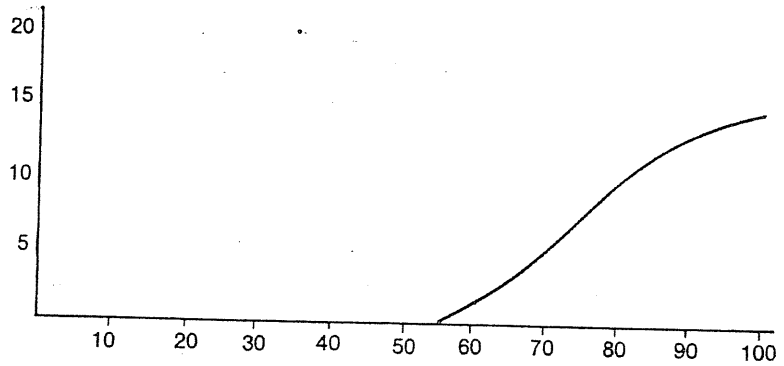
Achievement/attainment tests

This group can be further subdivided into class progress tests and (standardised) achievement tests.

Class progress tests

This book has been concerned chiefly with class progress tests, since these are the most widely used types of tests. Most teachers are, at some time or other, required to construct such tests. Each progress test situation is unique and can only be evaluated fully by the class teacher in the light of his or her knowledge of the students, the programme which they have been following, and the class teacher's own particular aims and goals. It is illogical to expect general purpose tests and books of tests to function as effectively as a test constructed specially for a particular situation: hence the purpose of this book in encouraging teachers to construct their own tests.

The progress test is designed to measure the extent to which the students have mastered the material taught in the classroom. It is based on the language programme which the class has been following and is just as important as an assessment of the teacher's own work as the student's own learning. Results obtained from the progress test enable the teacher to become more familiar with the work of each student and with the progress of the class in general. The class progress test is a teaching device, its backwash effect on teaching and motivation being important features. A good progress test should encourage the students to perform well in the target language and to gain additional confidence. Its aim is to stimulate learning and to reinforce what has been taught. Good performances act as a means of encouraging the students, and although poor performances may act as an incentive to more work, the progress test is chiefly concerned with allowing the students to show what they have mastered. Scores on it should thus be high (provided, of course, that progress has indeed been made). Whereas in standardised achievement and proficiency tests, a wide range of performance should be indicated, the progress test should show a cluster of scores around the top of the scale. Compare the following graph with the two graphs in 10.4.



Achievement tests

Achievement (or attainment) tests, though similar in a number of ways to progress tests, are far more formal tests and are intended to measure achievement on a larger scale. Most annual school examinations take the form of achievement tests; all public tests which are intended to show mastery of a particular syllabus are also achievement tests. These tests are based on what the students are presumed to have learnt – not necessarily on what they have actually learnt nor on what has actually been taught. Achievement tests frequently take the form of secondary school entrance tests and school certificate examinations; many are based on a published syllabus and exert a strong influence on the teaching in schools. Constructors of such tests rarely teach any of the students being tested (often an advantage *provided that* the test constructors are very familiar with the teaching and learning problems of the testees). Indeed, this is often a prerequisite before anyone can be appointed to any position of responsibility in connection with this type of test, though this principle obviously cannot always be applied to school examinations.

Several achievement tests are standardised: they are pre-tested, each item is analysed and revised where necessary, norms are established and comparisons made between performances of different students and different schools. Since such tests are administered year after year, it is possible to compare performances of students one year with those of students taking the test another year.

If the students have followed a structural approach to language learning, it is clearly unfair to administer a communicative achievement test at the end of their course. It is equally unfair to administer a structural-based test to those students who have followed a communicative approach to learning the target language. A good achievement test should reflect the particular approach to learning and teaching that has previously been adopted.

Proficiency tests

Whereas an achievement test looks back on what should have been learnt, the proficiency test looks forward, defining a student's language proficiency with reference to a particular task which he or she will be required to perform. Proficiency tests are in no way related to any syllabus or teaching programme; indeed, many proficiency tests are intended for students from several different schools, countries and even language backgrounds. The proficiency test is concerned simply with measuring a student's control of

the language in the light of what he or she will be expected to do with it in the future performance of a particular task. Does the student know enough English, for example, to follow a certain university or college course given in the medium of English? Does the student know enough English in order to function efficiently in a particular type of employment? The proficiency test is thus concerned with measuring not general attainment but specific skills in the light of the language demands made later on the student by a future course of study or job.

Aptitude tests

A language aptitude test (or prognostic test) is designed to measure the student's *probable* performance in a foreign language which he or she has not started to learn: i.e. it assesses aptitude for learning a language. Language learning aptitude is a complex matter, consisting of such factors as intelligence, age, motivation, memory, phonological sensitivity and sensitivity to grammatical patterning. The relative weighting given to these elements must depend on many factors and thus vary considerably from one individual to another. Some specialists in this field maintain that it is neither possible nor desirable to take an overall measurement of language aptitude; consequently aptitude is sometimes divided into various aspects according to the specific tasks for which a person is being trained: e.g. listening, interpreting, translating. Aptitude tests generally seek to predict the student's probable strengths and weaknesses in learning a foreign language by measuring performance in an artificial language. The ability to learn new phonemic distinctions and also to use language patterns in an unfamiliar but systematic manner is tested by means of the artificial language. Since few teachers are concerned with the complex field of aptitude testing, it is not necessary to go into further detail here.

Diagnostic tests

Although the term *diagnostic test* is widely used, few tests are constructed solely as diagnostic tests. Achievement and proficiency tests, however, are frequently used for diagnostic purposes: areas of difficulty are diagnosed in such tests so that appropriate remedial action can be taken later. Sections of tests which lend themselves particularly well to diagnostic purposes are phoneme discrimination tests, grammar and usage tests, and certain controlled writing tests. Clearly, weaknesses indicated in a test of vocabulary are not highly significant in themselves and can only be regarded as indicating general weaknesses. Similarly, many existing tests of reading comprehension are not very suitable for diagnostic purposes. Tests of writing and oral production can be used diagnostically provided that there is an appreciation of the limits to which such tests can be put. Since diagnosing strengths and weaknesses is such an important feature of progress tests and of teaching, the teacher should always be alert to every facet of achievement revealed in a class progress test.

Note that diagnostic testing is frequently carried out for groups of students rather than for individuals. If only one or two students make a particular error, the teacher will not pay too much attention. However, if several students in the group make a certain error, the teacher will note the error and plan appropriate remedial teaching.

Notes and references

1 The English Language Testing Service (ELTS), The British Council

11

Interpreting test scores

11.1 Frequency distribution

Marks awarded by counting the number of correct answers on a test script are known as raw marks. '15 marks out of a total of 20' may appear a high mark to some, but in fact the statement is virtually meaningless on its own. For example, the tasks set in the test may have been extremely simple and 15 may be the lowest mark in a particular group of scores.

TABLE 1

Testee	Mark
A	20
B	25
C	33
D	35
E	29
F	25
G	30
H	26
I	19
J	27
K	26
L	32
M	34
N	27
O	27
P	29
Q	25
R	23
S	30
T	26
U	22
V	23
W	33
X	26
Y	24
Z	26

TABLE 2

Testee	Mark	Rank
D	35	1
M	34	2
C	33	3.5 (or 3=)
W	33	3.5 (or 3=)
L	32	5
G	30	6.5 (or 6=)
S	30	6.5 (or 6=)
E	29	8.5 (or 8=)
P	29	8.5 (or 8=)
J	27	11 (or 10=)
N	27	11 (or 10=)
O	27	11 (or 10=)
H	26	15 (or 13=)
K	26	15 (or 13=)
T	26	15 (or 13=)
X	26	15 (or 13=)
Z	26	15 (or 13=)
B	25	19 (or 18=)
F	25	19 (or 18=)
Q	25	19 (or 18=)
Y	24	21
R	23	22.5 (or 22=)
V	23	22.5 (or 22=)
U	22	24
A	20	25
I	19	26

TABLE 3

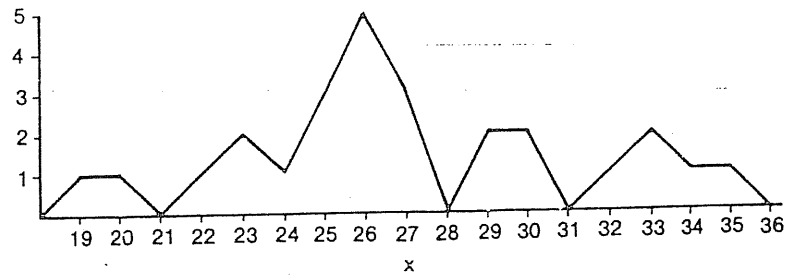
Mark	Tally	Frequency
40		
39		
38		
37		
36		
35	/	1
34	/	1
33	//	2
32	/	1
31		
30	//	2
29	//	2
28		
27	///	3
26	////	5
25	///	3
24	/	1
23	//	2
22	/	1
21		
20	/	1
19	/	1
18		
17		
16		
15	Total	26

Conversely, the test may have been extremely difficult, in which case 15 may well be a very high mark. Numbers still exert a strange and powerful influence on our society, but the shibboleth that 40 per cent should always represent a pass mark is nevertheless both surprising and disturbing.

The tables on the previous page contain the imaginary scores of a group of 26 students on a particular test consisting of 40 items. Table 1 conveys very little, but Table 2, containing the students' scores in order of merit, shows a little more. Table 3 contains a frequency distribution showing the number of students who obtained each mark awarded; the strokes on the left of the numbers (e.g. ////) are called *tallies* and are included simply to illustrate the method of counting the frequency of scores. Note that normally the frequency list would have been compiled without the need for Tables 1 and 2; consequently, as the range of highest and lowest marks would then not be known, all the *possible* scores would be listed and a record made of the number of students obtaining each score in the scale (as shown in the example).

Note that where ties occur in Table 2, two ways of rendering these are shown. The usual classroom practice is that shown in the parentheses. Where statistical work is to be done on the ranks, it is essential to record the average rank (e.g. testees J, N and O, each with the same mark, occupy places 10, 11 and 12 in the list, averaging 11).

The following frequency polygon illustrates the distribution of the scores:



11.2 Measures of central tendency

Mode

The *mode* refers to the score which most candidates obtained: in this case it is 26, as five testees have scored this mark.

Median

The *median* refers to the score gained by the middle candidate in the order of merit: in the case of the 26 students here (as in all cases involving even numbers of testees), there can obviously be no middle person and thus the score halfway between the lowest score in the top half and the highest score in the bottom half is taken as the median. The median score in this case is also 26.

Mean

The *mean* score of any test is the arithmetical average: i.e. the sum of the separate scores divided by the total number of testees. The mode, median, and mean are all measures of central tendency. The mean is the most efficient measure of central tendency, but it is not always appropriate.

In the following Table 4 and formula, note that the symbol x is used to denote the score, N the number of the testees, and m the mean. The

symbol f denotes the frequency with which a score occurs. The symbol Σ means *the sum of*.

TABLE 4		
x	f	fx
35 × 1		35
34 × 1		34
33 × 2		66
32 × 1		32
30 × 2		60
29 × 2		58
27 × 3		81
26 × 5		130
25 × 3		75
24 × 1		24
23 × 2		46
22 × 1		22
20 × 1		20
19 × 1		19
Total =		702 = Σfx

$$m = \frac{\Sigma fx}{N} = \frac{702}{26} = 27$$

Note that x = 702 is the total number of items which the group of 26 students got right between them. Dividing by N = 26 (as the formula states), this obviously gives the average.

It will be observed that in this particular case there is a fairly close correspondence between the mean (27) and the median (26). Such a close correspondence is not always common and has occurred in this case because the scores tend to cluster symmetrically around a central point.

11.3 Measures of dispersion

Whereas the previous section was concerned with measures of central tendency, this section is related to the range or spread of scores. The mean by itself enables us to describe an individual student's score by comparing it with the average set of scores obtained by a group, but it tells us nothing at all about the highest and lowest scores and the spread of marks.

Range

One simple way of measuring the spread of marks is based on the difference between the highest and lowest scores. Thus, if the highest score on a 50-item test is 43 and the lowest 21, the range is from 21 to 43: i.e. 22. If the highest score, however, is only 39 and the lowest 29, the range is 10. (Note that in both cases, the mean may be 32.) The range of the 26 scores given in Section 11.1 is: 35 − 19 = *16*.

Standard deviation

The standard deviation (s.d.) is another way of showing the spread of scores. It measures the degree to which the group of scores deviates from the mean; in other words, it shows how *all* the scores are spread out and thus gives a fuller description of test scores than the range, which simply describes the gap between the highest and lowest marks and ignores the information provided by all the remaining scores. Abbreviations used for the standard deviation are either s.d. or σ (the Greek letter sigma) or s.

One simple method of calculating s.d. is shown below:

$$\text{s.d.} = \sqrt{\frac{\Sigma d^2}{N}}$$

N is the number of scores and d the deviation of each score from the mean. Thus, working from the 26 previous results, we proceed to:

1. find out the amount by which each score deviates from the mean (d);
2. square each result (d^2);
3. total all the results (Σd^2);
4. divide the total by the number of testees ($\Sigma d^2/N$); and
5. find the square root of this result ($\sqrt{\Sigma d^2/N}$).

	Score	Mean Deviation (d)	Squared (d^2)	
(Step 1)	35 deviates from 27 by	8	(Step 2)	64
	34	7		49
	33	6		36
	33	6		36
	32	5		25
	30	3		9
	30	3		9
	29	2		4
	29	2		4
	27	0		0
	27	0		0
	27	0		0
	26	−1		1
	26	−1		1
	26	−1		1
	26	−1		1
	26	−1		1
	25	−2		4
	25	−2		4
	25	−2		4
	24	−3		9
	23	−4		16
	23	−4		16
	22	−5		25
	20	−7		49
	19	−8		64
	702	(Step 3)	Total	= 432

(Step 4) s.d. $= \sqrt{\frac{432}{26}}$

(Step 5) s.d. $= \sqrt{16.62}$

s.d. $= 4.077$

$= \underline{\underline{4.08}}$

Note: If deviations (d) are taken from the mean, their sum (taking account of the minus sign) is zero + 42 − 42 = 0. This affords a useful check on the calculations involved here.

A standard deviation of 4.08, for example, shows a smaller spread of scores than, say, a standard deviation of 8.96. If the aim of the test is simply to determine which students have mastered a particular programme

of work or are capable of carrying out certain tasks in the target language, a standard deviation of 4.08 or any other denoting a fairly narrow spread will be quite satisfactory provided it is associated with a high average score. However, if the test aims at measuring several levels of attainment and making fine distinctions within the group (as perhaps in a proficiency test), then a broad spread will be required.

Standard deviation is also useful for providing information concerning characteristics of different groups. If, for example, the standard deviation on a certain test is 4.08 for one class, but 8.96 on the same test for another class, then it can be inferred that the latter class is far more heterogeneous than the former.

11.4 Item analysis

Earlier careful consideration of objectives and the compilation of a table of test specifications were urged before the construction of any test was attempted. What is required now is a knowledge of how far those objectives have been achieved by a particular test. Unfortunately, too many teachers think that the test is finished once the raw marks have been obtained. But this is far from the case, for the results obtained from objective tests can be used to provide valuable information concerning:

- the performance of the students as a group, thus (in the case of class progress tests) informing the teacher about the effectiveness of the teaching;
- the performance of individual students; and
- the performance of each of the items comprising the test.

Information concerning the performance of the students as a whole and of individual students is very important for teaching purposes, especially as many test results can show not only the types of errors most frequently made but also the actual reasons for the errors being made. As shown in earlier chapters, the great merit of objective tests arises from the fact that they can provide an insight into the mental processes of the students by showing very clearly what choices have been made, thereby indicating definite lines on which remedial work can be given.

The performance of the test items, themselves, is of obvious importance in compiling future tests. Since a great deal of time and effort are usually spent on the construction of good objective items, most teachers and test constructors will be desirous of either using them again without further changes or else adapting them for future use. It is thus useful to identify those items which were answered correctly by the more able students taking the test and badly by the less able students. The identification of certain difficult items in the test, together with a knowledge of the performance of the individual distractors in multiple-choice items, can prove just as valuable in its implications for teaching as for testing.

All items should be examined from the point of view of (1) their difficulty level and (2) their level of discrimination.

Item difficulty

The *index of difficulty* (or *facility value*) of an item simply shows how easy or difficult the particular item proved in the test. The index of difficulty (FV) is generally expressed as the fraction (or percentage) of the students who answered the item correctly. It is calculated by using the formula:

$$FV = \frac{R}{N}$$

R represents the number of correct answers and N the number of students taking the test. Thus, if 21 out of 26 students tested obtained the correct answer for one of the items, that item would have an index of difficulty (or a facility value) of .77 or 77 per cent.

$$FV = \frac{20}{26} = .77$$

In this case, the particular item is a fairly easy one since 77 per cent of the students taking the test answered it correctly. Although an average facility value of .5 or 50 per cent may be desirable for many public achievement tests and for a few progress tests (depending on the purpose for which one is testing), the facility value of a large number of individual items will vary considerably. While aiming for test items with facility values falling between .4 and .6, many test constructors may be prepared in practice to accept items with facility values between .3 and .7. Clearly, however, a very easy item, on which 90 per cent of the testees obtain the correct answer, will not distinguish between above-average students and below-average students as well as an item which only 60 per cent of the testees answer correctly. On the other hand, the easy item will discriminate amongst a group of below-average students; in other words, one student with a low standard may show that he or she is better than another student with a low standard through being given the opportunity to answer an easy item. Similarly, a very difficult item, though failing to discriminate among most students, will certainly separate the good student from the very good student.

A further argument for including items covering a range of difficulty levels is that provided by motivation. While the inclusion of difficult items may be necessary in order to motivate the good student, the inclusion of very easy items will encourage and motivate the poor student. In any case, a few easy items can provide a 'lead-in' for the student – a device which may be necessary if the test is at all new or unfamiliar or if there are certain tensions surrounding the test situation.

Note that it is possible for a test consisting of items each with a facility value of approximately .5 to fail to discriminate at all between the good and the poor students. If, for example, half the items are answered correctly by the good students and incorrectly by the poor students while the remaining items are answered incorrectly by the good students but correctly by the poor students, then the items will work against one another and no discrimination will be possible. The chances of such an extreme situation occurring are very remote indeed; it is highly probable, however, that at least one or two items in a test will work against one another in this way.

Item discrimination

The discrimination index of an item indicates the extent to which the item discriminates between the testees, separating the more able testees from the less able. The index of discrimination (D) tells us whether those students who performed well on the whole test tended to do well or badly on each item in the test. It is presupposed that the total score on the test is a valid measure of the student's ability (i.e the good student tends to do well on the test as a whole and the poor student badly). On this basis, the score on the whole test is accepted as the criterion measure, and it thus becomes possible to separate the 'good' students from the 'bad' ones in performances on individual items. If the 'good' students tend to do well on

an item (as shown by many of them doing so – a frequency measure) and the 'poor' students badly on the same item, then the item is a good one because it distinguishes the 'good from the 'bad' in the same way as the total test score. This is the argument underlying the index of discrimination.

There are various methods of obtaining the index of discrimination: all involve a comparison of those students who performed well on the whole test and those who performed poorly on the whole test. However, while it is statistically most efficient to compare the top 27½ per cent with the bottom 27½ per cent, it is enough for most purposes to divide small samples (e.g. class scores on a progress test) into halves or thirds. For most classroom purposes, the following procedure is recommended.

1 Arrange the scripts in rank order of total score and divide into two groups of equal size (i.e. the top half and the bottom half). If there is an odd number of scripts, dispense with one script chosen at random.
2 Count the number of those candidates in the upper group answering the first item correctly; then count the number of lower-group candidates answering the item correctly.
3 Subtract the number of correct answers in the lower group from the number of correct answers in the upper group: i.e. find the difference in the proportion passing in the upper group and the proportion passing in the lower group.
4 Divide this difference by the total number of candidates in one group:

$$D = \frac{\text{Correct U} - \text{Correct L}}{n}$$

(D = Discrimination index; n = Number of candidates in one group*; U = Upper half and L = Lower half. The index D is thus the difference between the proportion passing the item in U and L.)
5 Proceed in this manner for each item.

The following item, which was taken from a test administered to 40 students, produced the results shown:

I left Tokyo . . . . . Friday morning.
A. in (B.) on C. at D. by

$$D = \frac{15 - 6}{20} = \frac{9}{20} = .45$$

Such an item with a discrimination index of .45 functions fairly effectively, although clearly it does not discriminate as well as an item with an index of .6 or .7. Discrimination indices can range from + 1 (= an item which discriminates perfectly – i.e. it shows perfect *correlation* with the testees' results on the whole test) through 0 (= an item which does not discriminate in any way at all) to −1 (= an item which discriminates in entirely the wrong way). Thus, for example, if all 20 students in the upper group answered a certain item correctly and all 20 students in the lower group got the wrong answer, the item would have an index of discrimination of 1.0.

*The reader should carefully distinguish between n(= the number of candidates in either the U or L group) and N (= the number in the whole group) as used previously. Obviously n = 1/2 N.

If, on the other hand, only 10 students in the upper group answered it correctly and furthermore 10 students in the lower group also got correct answers, the discriminaton index would be 0. However, if none of the 20 students in the upper group got a correct answer and all the 20 students in the lower group answered it correctly, the item would have a negative discrimination, shown by −1.0. It is highly inadvisable to use again, or even to attempt to amend, any item showing negative discrimination. Inspection of such an item usually shows something radically wrong with it.

Again, working from actual test results, we shall now look at the performance of three items. The first of the following items has a high index of discrimination; the second is a poor item with a low discrimination index; and the third example is given as an illustration of a poor item with negative discrimination.

1 High discrimination index:

NEARLY When ..... Jim ..... crossed ..... the road, he ..... ran into a car.

$$D = \frac{18 - 3}{20} = \frac{15}{20} = .75 \qquad FV = \frac{21}{40} = \underline{0.525}$$

(The item is at the right level of difficulty and discriminates well.)

2 Low discrimination index:

If you ..... the bell, the door would have been opened.

A. would ring
(B.) had rung
C. would have rung
D. were ringing

$$D = \frac{3 - 0}{20} = \underline{.15} \qquad FV = \frac{3}{40} = \underline{.075}$$

(In this case, the item discriminates poorly because it is too difficult for everyone, both 'good' and 'bad'.)

3 Negative discrimination index:

I don't think anybody has seen him.

(A.) Yes, someone has.
(B.) Yes, no one has.
C. Yes, none has.
D. Yes, anyone has.

$$D = \frac{4 - 6}{20} = \frac{-2}{20} = \underline{-.10} \qquad FV = \frac{10}{40} = \underline{0.25}$$

(This item is too difficult and discriminates in the wrong direction.)

What has gone wrong with the third item above? Even at this stage and without counting the number of candidates who chose each of the options, it is evident that the item was a trick item: in other words, the item was far too 'clever'. It is even conceivable that many native speakers would select option B in preference to the correct option A. Items like this all too often escape the attention of the test writer until an item analysis actually focuses attention on them. (This is one excellent reason for conducting an item analysis.)

Note that items with a very high facility value fail to discriminate and thus generally show a low discrimination index. The particular group of

students who were given the following item had obviously mastered the use of *for* and *since* following the present perfect continuous tense:

He's been living in Berlin ..... 1975.

$$D = \frac{19 - 19}{20} = \underline{0} \qquad FV = \frac{38}{40} = \underline{0.95}$$

(The item is extremely easy for the testees and has zero discrimination.)

Item difficulty and discrimination

Facility values and discrimination indices are usually recorded together in tabular form and calculated by similar procedures. Note again the formulae used:

$$FV = \frac{\text{Correct U} + \text{Correct L}}{2n} \qquad \left(\text{or } FV = \frac{R}{N} \right)$$

$$D = \frac{\text{Correct U} - \text{Correct L}}{n}$$

The following table, compiled from the results of the test referred to in the preceding paragraphs, shows how these measures are recorded.

Item	U	L	U+L	FV	U−L	D
1	19	19	38	.95	0	0
2	13	16	29	.73	−3	−.15
3	20	12	32	.80	8	.40
4	18	3	21	.53	15	.75
5	15	6	21	.53	9	.45
6	16	15	31	.77	1	.05
7	17	8	25	.62	9	.45
8	13	4	17	.42	9	.45
9	4	6	10	.25	−2	−.10
10	10	4	14	.35	6	.30
11	18	13	31	.78	5	.25
12	12	2	14	.35	10	.50
13	14	6	20	.50	8	.40
14	5	1	6	.15	4	.20
15	7	1	8	.20	6	.30
16	3	0	3	.08	3	.15
Etc.						

Items showing a discrimination index of below .30 are of doubtful use since they fail to discriminate effectively. Thus, on the results listed in the table above, only items 3, 4, 5, 7, 8, 10, 12, 13 and 15 could be safely used in future tests without being rewritten. However, many test writers would keep item 1 simply as a lead-in to put the students at ease.

Extended answer analysis

It will often be important to scrutinise items in greater detail, particularly in those cases where items have not performed as expected. We shall want to know not only why these items have not performed according to expectations but also why certain testees have failed to answer a particular item correctly. Such tasks are reasonably simple and straightforward to perform if the multiple-choice technique has been used in the test.

In order to carry out a full item analysis, or an extended answer analysis, a record should be made of the different options chosen by each student in the upper group and then the various options selected by the lower group.

If I were rich, I . . . . . work.

A. shan't B. won't (C.) wouldn't D. didn't

	U	L	U+L
A.	1	4	5
B.	2	5	7
C.	14	4	18
D.	3	7	10
	(20)	(20)	(40)

$$FV = \frac{U+L}{2n} = \frac{18}{40} = .45$$

$$D = \frac{U-L}{n} = \frac{10}{20} = .50$$

The item has a facility value of .45 and a discrimination index of .50 and appears to have functioned efficiently: the distractors attract the poorer students but not the better ones.

The performance of the following item with a low discrimination index is of particular interest:

Mr Watson wants to meet a friend in Singapore this year.
He . . . . . him for ten years.

A. knew B. had known C. knows (D.) has known

	U	L	U+L
A.	7	3	10
B.	4	3	7
C.	1	9	10
D.	8	5	13
	(20)	(20)	(40)

FV = .325
D = .15

While distractor C appears to be performing well, it is clear that distractors A and B are attracting the wrong candidates (i.e. the better ones). On closer scrutiny, it will be found that both of these options may be correct in certain contexts: for example, a student may envisage a situation in which Mr Watson is going to visit a friend whom he had known for ten years in England but who now lives in Singapore, e.g.

He knew him (well) for ten years (while he lived in England).

The same justification applies for option B.

The next item should have functioned efficiently but failed to do so: an examination of the testees' answers leads us to guess that possibly many had been taught to use the past perfect tense to indicate an action in the past taking place before another action in the past. Thus, while the results obtained from the previous item reflect on the item itself, the results here *probably* reflect on the teaching:

John F. Kennedy . . . . . born in 1917 and died in 1963.

A. is B. has been (C.) was D. had been

	U	L	U+L	
A.	0	2	2	
B.	0	3	3	FV = .625
C.	13	12	25	D = .05
D.	7	3	10	
	(20)	(20)	(40)	

In this case, the item might be used again with another group of students, although distractors A and B do not appear to be pulling much weight.

Distractor D in the following example is ineffective and clearly needs to be replaced by a much stronger distractor:

He complained that he . . . . . the same bad film the night before.

(A.) had seen B. was seeing C. has seen D. would see

	U	L	U+L	
A.	14	8	22	
B.	4	7	11	FV = .55
C.	2	5	7	D = .30
D.	0	0	0	
	(20)	(20)	(40)	

Similarly, the level of difficulty of distractors C and D in the following item is far too low: a full item analysis suggests only too strongly that they have been added simply to complete the number of options required.

Wasn't that your father over there?

A. Yes, he was. C. Yes, was he.
(B.) Yes, it was. D. Yes, was it.

	U	L	U+L	
A.	7	13	20	
B.	13	7	20	FV = .50
C.	0	0	0	D = .30
D.	0	0	0	
	(20)	(20)	(40)	

The item could be made slightly more difficult and thus improved by replacing distractor C by *Yes, he wasn't* and D by *Yes, it wasn't*. The item is still imperfect, but the difficulty level of the distractors will probably correspond more closely to the level of attainment being tested.

The purpose of obtaining test statistics is to assist interpretation of item and test results in a way which is meaningful and significant. Provided that such statistics lead the teacher or test constructor to focus once again on the *content* of the test, then item analysis is an extremely valuable exercise. Only when test constructors misapply statistics or become uncritically dominated by statistical procedures does item analysis begin to exert a harmful influence on learning and teaching. In the final analysis, the teacher should be prepared to sacrifice both reliability and discrimination to a limited extent in order to include in the test certain items which he or she regards as having a good 'educational' influence on

the students if, for example, their exclusion might lead to neglect in teaching what such items test.

11.5 Moderating

The importance of moderating classroom tests as well as public examinations cannot be stressed too greatly. No matter how experienced test writers are, they are usually so deeply involved in their work that they become incapable of standing back and viewing the items with any real degree of objectivity. There are bound to be many blind-spots in tests, especially in the field of objective testing, where the items sometimes contain only the minimum of context.

It is essential, therefore, that the test writer submits the test for moderation to a colleague or, preferably, to a number of colleagues. Achievement and proficiency tests of English administered to a large test population are generally moderated by a board consisting of linguists, language teachers, a psychologist, a statistician, etc. The purpose of such a board is to scrutinise as closely as possible not only each item comprising the test but also the test as a whole, so that the most appropriate and efficient measuring instrument is produced for the particular purpose at hand. In these cases, moderation is also frequently concerned with the scoring of the test and with the evaluation of the test results.

The class teacher does not have at his or her disposal all the facilities which the professional test writer has. Indeed, it is often all too tempting for the teacher to construct a test without showing it to anyone, especially if the teacher has had previous training or experience in constructing examinations of the more traditional type. Unfortunately, few teachers realise the importance of making a systematic analysis of the elements and skills they are trying to test and, instead of compiling a list of test specifications, tend to select testable points at random from coursebooks and readers. Weaknesses of tests constructed in this manner are brought to light in the process of moderation. Moreover, because there is generally more than one way of looking at something, it is incredibly easy (and common) to construct multiple-choice items containing more than one correct option. In addition, the short contexts of many objective items encourage ambiguity, a feature which can pass by the individual unnoticed. To the moderator, some items in a test may appear far too difficult or else far too easy, containing implausible distractors; others may contain unsuspected clues. Only by moderation can such faults be brought to the attention of the test writer.

In those cases where the teacher of English is working on his or her own in a school, assistance in moderation from a friend, a spouse, or an older student will prove beneficial. It is simply impossible for any single individual to construct good test items without help from another person.

11.6 Item cards and banks

As must be very clear at this stage, the construction of objective tests necessitates taking a great deal of time and trouble. Although the scoring of such tests is simple and straightforward, further effort is then spent on the evaluation of each item and on improving those items which do not perform satisfactorily. It seems somewhat illogical, therefore, to dispense with test items once they have appeared in a test.

The best way of recording and storing items (together with any relevant information) is by means of small cards. Only one item is entered on each card; on the reverse side of the card information derived from an item analysis is recorded: e.g. the facility value (FV), the Index of

Discrimination (D), and an extended answers analysis (if carried out). After being arranged according to the element or skill which they are intended to test, the items on the separate cards are grouped according to difficulty level, the particular area tested, etc. It is an easy task to arrange them for quick reference according to whatever system is desired. Furthermore, the cards can be rearranged at any later date.

Although it will obviously take considerabie time to build up an item bank consisting of a few hundred items, such an item bank will prove of enormous value and will save the teacher a great deal of time and trouble later. The same items can be used many times again, the order of the items (or options within each item) being changed each time. If there is concern about test security or if there is any other reason indicating the need for new items, many of the existing items can be rewritten. In such cases, the same options are generally kept, but the context is changed so that one of the distractors now becomes the correct option. Multiple-choice items testing most areas of the various language elements and skills can be rewritten in this way, e.g.

(Grammar)

I hope you ..... us your secret soon.
A. told B. will tell C. have told D. would tell

→ I wish you ..... us your secret soon.
A. told B. will tell C. have told D. would tell

(Vocabulary)

Are you going to wear your best ..... for the party?
A. clothes B. clothing C. cloths D. clothings

→ What kind of ..... is your new suit made of?
A. clothes B. clothing C. cloth D. clothings

(Phoneme discrimination)

beat bit beat
→ beat beat bit

(Listening comprehension)

Student hears: Why are you going home?
Student reads:
A. At six o'clock.
B. Yes, I am.
C. To help my mother.
D. By bus.

→ Student hears: How are you going to David's?
Student reads:
A. At six o'clock.
B. Yes, I am.
C. To help him.
D. By bus.

(Reading comprehension/ vocabulary) → Two-thirds of the country's (fuel, endeavour, industry, energy) comes from imported oil, while the remaining one-third comes from coal. Moreover, soon the country will have its first nuclear power station.

Two-thirds of the country's (fuel, endeavour, industry, power) takes the form of imported oil, while the remaining one-third is coal. However, everyone in the country was made to realise the importance of coal during the recent miners' strike, when many factories were forced to close down.

Items rewritten in this way become new items, and thus it will be necessary to collect facility values and discrimination indices again.

Such examples serve to show ways of making maximum use of the various types of test items which have been constructed, administered and evaluated. In any case, however, the effort spent on constructing tests of English as a second or foreign language is never wasted since the insights provided into language behaviour as well as into language learning and teaching will always be invaluable in any situation connected with either teaching or testing.

Selected bibliography

Closed tests and examinations in English as a second/foreign language

Associated Examining Board (AEB)
Test in English for Educational Purposes (TEEP)

Association of Recognized English Language Schools Examination Trust (AET)
ARELS Oral Examinations: Preliminary, Higher Certificate, Diploma

British Council/University of Cambridge Local Examinations Syndicate
English Language Testing Service

City and Guilds of London Institute
Communication in Technical English

Educational Testing Service (Princeton, New Jersey, USA)
Test of English as a Foreign Language (TOEFL)

English Language Teaching Development Unit (ELTDU)
Stages of Attainment Scale and Test Battery

English Speaking Board (International) Limited (ESB)
Oral Assessments in Spoken English as an Acquired Language

General Medical Council (GMC)
The PLAB Test

Institute of Linguists Educational Trust
Examinations in English as a Foreign Language: Levels 1, 2, 3 and 4
Certificates in English as a Foreign Language: Preliminary, Grade I and Grade II
Diplomas in English as a Foreign Language: Intermediate Diploma and Final Diploma

Joint Matriculation Board (JMB)
Test in English (Overseas)

London Chamber of Commerce and Industry (LCCI)
English for Commerce: Elementary Stage, Intermediate Stage, and Higher Stage
Spoken English for Industry and Commerce (SEFIC)

North West Regional Examinations Board/North Western Regional Advisory Council for Further Education
English as a Second Language: 2 versions/levels

Oxford Delegacy of Local Examinations (Oxford)
University of Oxford Delegacy's Examinations in English as a Foreign Language: Preliminary Level Certificate and Higher Level Certificate

Pitman Examinations Institute (PEI)
English as a Foreign Language (Syllabus L and Syllabus C) and English Language (NCE): Elementary, Intermediate, Higher Intermediate and Advanced

Royal Society of Arts Examinations Board (RSA)
English as a Foreign Language: Stage I, Stage II and Stage III
Communicative Use of English as a Foreign Language: Basic, Intermediate and Advanced

Trinity College, London
Spoken English as a Foreign or Second Language: 12 grades
Written English (Intermediate)

University of Cambridge Local Examinations Syndicate
Preliminary English Test (PET)
First Certificate in English (FCE)
Certificate of Proficiency in English (CPE)
Diploma of English Studies (DES)
School Certificate and GCE (Overseas Centres): English Language

University of London
O Level English Language Syllabus B

Books and articles

Aitken, K G 1977 Using cloze procedure as an overall language proficiency test. *TESOL Quarterly* **11**(1): 59–67

Aitken, K G 1979 Techniques for assessing listening comprehension in second languages. *Audio-Visual Language Journal* **17**: 175–81

Alderson, J C 1978 A study of the cloze procedure with native and non-native speakers of English. University of Edinburgh PhD thesis

Alderson, J C and Hughes, A (eds.) 1981 Issues in Language Testing. *ELT Documents III*. British Council

Allen J P B and Davies, A (eds.) 1977 Testing and experimental methods. *Edinburgh Course in Applied Linguistics* vol 4. Oxford University Press

Anderson, J 1971 A technique for measuring reading comprehension and readability. *English Language Teaching Journal* **25**(2): 178–82

Anderson, J 1976 *Psycholinguistic experiments in foreign language testing*. University of Queensland Press.

Beardsmore, H B 1974 Testing oral fluency. *IRAL* **12**(4): 317–26

Bensoussan, M 1983 Dictionaries and tests of EFL reading comprehension. *English Language Teaching Journal* **37**(4): 341–5

Brown, G 1977 *Listening to spoken English*. Longman

Brumfit, C J 1984 *Communicative Methodology in Language Teaching*. Cambridge University Press

Burstall, C 1969 The main stages in the development of language tests. In Stern, H H (ed.) *Languages and the Young School Child*. Oxford University Press, 193–9

Burt, M K and Kiparsky, C 1972 *The Gooficon: a repair manual for English*. Newbury House, Rowley, Massachusetts

Canale, M and Swain, M 1980 Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics* **1**: 1–47

Carroll, B J 1980 *Testing Communicative Performance*. Pergamon

Chen, Z and Henning, G 1985 Linguistic and cultural bias in language proficiency tests. *Language Testing* **2**(2): 180–191

Clark J L D 1972 *Foreign Language Testing: Theory and Practice*. Centre for Curriculum Development, Philadelphia, Pennsylvania

Clark, J L D (ed.) 1978 *Direct Testing of Speaking Proficiency: Theory and Practice*. Educational Testing Service, Princeton

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

Crocker, A C 1969 *Statistics for the Teacher (or How to Put Figures in their Place)*. Penguin

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

Davies, A 1978 Language Testing (Survey Articles Nos. 3 and 4). In Kinsella, V (ed.) *Language Teaching and Linguistics Abstracts* vol. 2: 127–59. Cambridge University Press

Davies, S and West, R 1984 *The Pitman Guide to English Language Examinations* 2nd edn. Pitman

Douglas, D 1978 Gain in reading proficiency in English as a Foreign Language measured by three cloze scoring methods. *Journal of Research in Reading* **1**(1): 67–73

Ellis, R 1984 Communication strategies and the evaluation of communicative performance. *English Language Teaching Journal* **38**(1): 39–44

Finocchiaro, M and Sako, S 1983 *Foreign Language Testing: A Practical Approach*. Regents, New York

Fok, A, Lord, R, Low, G, T'sou B K, and Lee, Y P 1981 *Working Papers in Linguistics and Language Teaching*. Special Issue on Language Testing, No. 4, Language Centre, University of Hong Kong

Gannon, P 1985 *Assessing Writing: principles and practice of marking written English*. Edward Arnold

Geoghegan, G 1983 *Language problems of non-native speakers of English at Cambridge University*. Bell Educational Trust, Cambridge

Green, J A 1975 *Teacher-Made Tests* 2nd edn. Harper and Row, New York

Hale, G A, Stansfield, C W, and Duran, R P 1984 *Summaries of studies involving the Test of English as a Foreign Language, 1963–1982*. Educational Testing Service, Princeton, New Jersey

Hanania, E and Shikhani, M 1986 Interrelationships Among Three Tests of Language Proficiency: Standardized ESL, Cloze, and Writing. *TESOL Quarterly* **20**(1): 97–109

Harris, D P 1969 *Testing English as a Second Language*. McGraw-Hill, New York

Harrison, A 1983 *A Language Testing Handbook*. Macmillan

Heaton, J B (ed.) 1982 *Language Testing*. Modern English Publications

Hendrickson, J (ed.) Error Analysis and Error Correction in Language Teaching. *RELC Occasional Papers* **10**, SEAMEO Regional Language Centre, Singapore

Henning, G H *et. al.* 1981 Comprehensive Assessment of Language Proficiency and Achievement Among Learners of English as a Foreign Language. *TESOL Quarterly* **15**(4): 457–66

Hughes, A and Lascaratou, C 1982 Competing criteria for error gravity. *English Language Teaching Journal* **63**(3): 175–82

Hughes, A and Porter, D (eds.) 1983 *Current Developments in Language Testing*. Academic Press

Hughes, A and Porter, D (eds.) *Language Testing*. (Journal published by Edward Arnold in June and December each year.)

Ibe, M D 1975 A comparison of cloze and multiple-choice tests for measuring the English reading comprehension of southeast Asian teachers of English. *RELC Journal* **6**(2): 24–32. SEAMEO Regional Language Centre, Singapore

Jones, R L and Spolsky, B (eds.) 1975 *Testing Language Proficiency*. Center for Applied Linguistics, Arlington, Virginia

Lado, R 1961, 1964 *Language Testing: the Construction and Use of Foreign Language Tests*. Longman

Lee, Y P and Low, G D 1981 *Classifying tests of language use*. Paper presented at 6th AILA World Congress, Lund, Sweden

Lee, Y P, Fok, C Y Y, Lord, R, and Low, G 1982 *New Directions in Language Testing*. Pergamon, Hong Kong

Lukmani, Y 1972 The communicational testing of reading. *English Language Teaching Journal* **36**(4): 217–25

Moller, A 1975 Validity in Proficiency Testing, *ELT Documents* **3**: 5–18. British Council

Morrow, K E 1977 *Techniques of Evaluation for a Notional Syllabus*. Centre for Applied Language Studies, University of Reading (for the Royal Society of Arts)

Morrow, K E 1979 Communicative Language Testing: revolution or evolution. In Brumfit, C J and Johnson, K J (eds.) *The Communicative Approach to Language Teaching*. Oxford University Press

Munby, J L 1978 *Communicative Syllabus Design*. Cambridge University Press

Oller, J W 1972 Cloze tests of second language proficiency and what they measure. *Language Learning* **23**(1): 105–18

Oller, J W 1979 *Language Tests at School*. Longman

Oller, J W and Perkins, K 1978 *Language in Education: testing the tests*. Newbury House, Rowley, Massachusetts

Oller, J W and Streiff, V 1975 Dictation: a test of grammar-based expectancies. *English Language Teaching Journal* **30**(1): 25–36

Palmer, A S 1981 Testing communication. *IRAL* **10**: 35–45

Palmer, A S 1981 Measures of achievement, communication, incorporation, and integration for two classes of formal EFL learners. *RELC Journal* **12**(1): 37–61

Palmer, L and Spolsky, B (eds.) 1975 *Papers on Language Testing, 1967–1974*. TESOL, Washington, D. C.

Perkins, K 1980 Using Objective Methods of Attained Writing Proficiency to Discriminate Among Holistic Evaluations. *TESOL Quarterly* **14**(1): 61–69

Perren, G E (ed.) 1977 *Foreign Language Testing: specialised bibliography*. Centre for Information on Language Teaching and Research

Portal, M (ed.) 1986 *Innovations in Language Testing*. NFER-Nelson

Rea, P M 1978 Assessing language as communication: *MALS Journal*. New series, No. 3, Department of English, University of Birmingham

Read, J A S (ed.) 1981 *Directions in Language Testing, RELC Anthology, Series 9*. SEAMEO Regional Language Centre, Singapore

Richards, J C 1985 *The Context of Language Teaching*. Cambridge University Press

Rivera, C 1984 *Communicative Competence Approaches to Language Proficiency Assessment: Research and Application*. Multilingual Matters, Clevedon

Rivers, W M 1968 *Teaching foreign-language skills*. University of Chicago Press

Rivers, W M and Temperley, M S 1978 *A Practical Guide to the Teaching of English as a Second or Foreign Language*. Oxford University Press, New York

Schulz, R A 1977 Discrete-Point versus Simulated Communication Testing in Foreign Languages. *Modern Language Journal* **61**(3): 91–101

Simmonds, P 1985 A survey of English language examinations. *English Language Teaching Journal*, **39**(1): 33–42

Spolsky, B 1985 What does it mean to know how to use a language? An essay on the theoretical basis of language testing. *Language Testing* **2**(2): 180–191

Spolsky, B with Murphy, P, Holm, W, and Ferrel, A 1972 Three Functional Tests of Oral Proficiency. *TESOL Quarterly* **6**(3): 221–35

Stubbs, J B and Tucker, G R 1974 The Cloze Test as a Measure of English Proficiency. *Modern Language Journal* **58**(5/6): 239–41

Tomiyana, M 1980 Grammatical errors and communication breakdown. *TESOL Quarterly* **14**(1): 71–9

Upshur, J A 1971 Objective evaluation of oral proficiency in the ESOL classroom. *TESOL Quarterly* **5**: 47–60

Upshur, J A and Fata, J 1968 Problems in foreign language testing. *Language Learning* Special Issue, No. 3

Valette, R M 1977 *Modern Language Testing* 2nd edn. Harcourt Brace Jovanovich, New York

Valette, R M and Disick, R S 1972 *Modern Language Performance Objectives and Individualisation*. Harcourt Brace Jovanovich, New York

Index

Achievement tests 172
Addition items 50
Administration 167–168
Analytic marking 148
Aptitude items 173
Attainment tests 171

Backwash 170–171
Banding
— compositions 145–146
— oral interviews 98–100
Broken sentence items 49

Central tendency 175–176
Changing words 48–49
Classification of tests 15
Classroom tests 6
Cloze procedure 16–17, 131–133
Combination items 50
Communicative approach 19–24
Completion items
— grammar 42–46
— reading comprehension 124–129
— sentences, texts 156–157
— spelling 151–152
— vocabulary 62–63
Composition
— general 136–150
— setting 138–143
— titles 137–138
Concurrent validity 161–162
Connectives 156
Construct validity 161
Content validity 160–161
Controlled writing 154–158
Conversations 90–92
Cursory reading 133–134

Definitions 62
Diagnostic testing 6, 173
Dialogues 69–71, 90–92
Dictation 17–18, 151
Difficulty index 178–179
Discrimination 165, 179–182
Dispersion 176–178

Empirical validity 161–162
Error, test margin of 166
Error recognition items 39–40, 152
Error-count marking 148–149
Errors 149–150
Essay–translation approach 15
Evaluation 7
Extended answer analysis 182–185
Extensive reading 106

Face validity 159–160
Facility value 178–179
Fragmented sentences 155
Frequency distribution 174–175

Grading compositions 144–149
Grammar 9, 34–50
Group discussion 102–104

Impression marking 147–148
Impure items 29
Index of difficulty 178–179
Instructions 168–170
Integrative approach 16
Intensive reading 106
Interpretation of scores 174–188
Intonation 68–69
Items
— analysis 178–185
— cards and banks 185–187
— difficulty 178–179
— discrimination 179–182
(*For types of items, see individual entries, e.g. addition, combination, multiple-choice*)

Judgement skills 135

Language areas 9
Language elements 10–11
Language skills 8, 10–11
Lectures 82–87
Length of texts 118
Linking sentences 15
Listening comprehension 64–87

Mark/re-mark reliability 162
Marking
— analytic method 148
— compositions 144–149
— error-count method 148–149
— impression method 147–148
— mechanical accuracy method 148–149
— treatment of errors 149–150
Matching items
— reading 107–113
— vocabulary 58–60
Mean 175–176
Mechanical accuracy marking 148–149
Median 175
Mistakes 7, 149–150
Mode 175
Model paragraphs 154
Moderating 186–187
Motivation 7
Multiple marking 147–148
Multiple-choice items
— construction 27–40
— correct option 37–38
— distractors 32–33
— error recognition 39–40
— grammar 9
— listening 66–71
— reading 116–124
— spelling 151
— stem 36–37
— vocabulary 52–28
— writing 152–153

Objective testing 25–27
Open-ended items 133
Oral production 88–104
Oral interview 96–102

Pairing and matching items 49–50
Parallel test forms 163
Performance levels 21–23
Phoneme discrimination 65–68
Phonology 9
Pictures
— listening comprehension 71–82
— matching 110, 112
— speaking 92–96
— writing 142–143
Practicability 167–168
Predictive validity 161–162
Problem solving 102–104
Production 11

Proficiency tests 172–173
Profile reporting 19–23, 163
Progress tests 171–172
Psychometric approach 16
Punctuation 135, 150–151

Qualitative judgements 20–22

Range 176
Rating scale
— compositions 145–146
— oral interviews 98–100
Reading
— extracts (for writing) 155–156
— reading aloud 89–90
— skills 105–106
— test specifications 22–23
Rearrangement items
— grammar 41–42
— reading 129–131
— vocabulary 61
Recognition 11
Redundancy 64
Register 153–154
Reliability 12, 162–165
— listening comprehension 65
— mark/re-mark 162
— parallel test forms 163
— profile reporting 163
— split half-method 163–164
— test/re-test 162
— v validity 164–165
Role playing 102–104
Rubrics 168–170

Sampling 12, 51–52, 118
Scanning 133–134
Scores 174–188
Scoring
— oral interviews 98–100
Sentence matching 107–110
Sets 58
Skimming 133–134
Speaking 21–23, 88–104
Specifications 13
Spelling 135, 151–152
Split-half method 163–164
Spoken language 64–65
Spread of scores 166–167
Standard deviation 176–178
Standards 7
Statements 69–71
Stress 68–69
Structuralist approach 15
Style 152–153
Subjective testing 25–26
Synonyms 61

Talks 82–87, 102
Teaching 5, 170–171
Test types 106–107, 171–173
Test/re-test reliability 162
Transformation items 46–48
Translation 18–19
Traps 14
True/false tests 113–116

Usage 34–50

Validity 159–162
— concurrent 161
— construct 161
— content 160–161
— empirical 161–162
— face 159–160
— predictive 161
— v reliability 164–165
Visuals
— listening comprehension 71–82
— matching 110, 112
— speaking 92–96
— writing 142–143
Vocabulary 9, 51–63

Word formation 61
Word matching 107
Writing 135–158
— levels 136
— tasks 136